POPULATION GENOMICS OF HUMAN POLYMORPHIC TRANSPOSABLE ELEMENTS

A Dissertation
Presented to
The Academic Faculty

by

Lavanya Rishishwar

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics in the
School of Biological Sciences

Georgia Institute of Technology
December 2016

POPULATION GENOMICS OF HUMAN POLYMORPHIC TRANSPOSABLE ELEMENTS

Approved by:

Dr. I. King Jordan, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. John F. McDonald
School of Biological Sciences
Georgia Institute of Technology

Dr. Greg Gibson
School of Biological Sciences
Georgia Institute of Technology

Dr. Soojin V. Yi
School of Biological Sciences
Georgia Institute of Technology

Dr. Leonardo Mariño-Ramírez
National Center for Biotechnology Information
National Institutes of Health

Date Approved: November 7, 2016

To my family and friends

ACKNOWLEDGEMENTS

I am deeply grateful to my advisor Dr. I. King Jordan for his constant support and guidance throughout my time here as a student, working with him and his lab. I will always be appreciative of his patience and the time he took teach me many fundamental things of science and scientific communication. His passion of science is contagious and has left a large impression on me. He is the finest mentor I've ever had, a great friend and a father-like figure to me. Under his tutelage, I've become a better researcher, better teacher and, most of all, a better person.

I had the pleasure of having Dr. Greg Gibson, Dr. John McDonald, Dr. Soojin Yi and Dr. Leonardo Mariño-Ramírez as the most supportive committee members I could've ever had. I would also like to thank Dr. Joseph Lachance for all the helpful discussions and insights. Their constant encouragement throughout this program was crucial for my success and development.

I am grateful to my friends and colleagues from the Jordan Lab, Dr. Andrew Conley, Dr. Jianrong Wang, Dr. Daudi Jjingo, Dr. Lee Katz, Eishita Tyagi and Lu Wang, who motivated and inspired me to do amazing things. I want to thank Emily Norris for being a great friend and giving me her unconditional support over the past few years. I am extremely thankful of her time with all of the helpful discussions over my numerous papers. I am also obliged to Dr. Vinay Mittal, Piyush Ranjan and Siddharth Choudhary for the fun and crazy moments that helped me get through grad school in a stress free manner.

I am very appreciative of Dr. Bala Swaminathan for trusting me and giving me the opportunity to lead the Applied Bioinformatics Laboratory (ABiL) and broaden my horizons. I also thank Michael Astwood for trusting me and constantly pushing me to believe in myself. You have been a great friend and colleague.

I must also thank Troy Hilley, for being patient and helping me out with all the IT related problems, and Lisa Redding, for always being responsive and helping me out with all the program related requirements. Their support saved me a lot of time and frustration over the years.

I am very much appreciative of my friends in Atlanta: Suzanna Kim, Angela Peña, Juliana Soto, Kizee Etienne, Linh Chau, Chinar Patil, Laurel Jenkins, Kawther Abdilleh, Camila Medrano and Aroon Chande for the memorable times we have had. Your company helped me to feel that Atlanta was my home.

Last, but not the least, I couldn't have done all of this without the constant love and support from my family, my brother Kshitij Rishishwar, my father Sanjay Rishishwar, my mother Dr. Poonam Rishishwar and my grandfather Dr. Krishna Mohan. This was their dream and my sole reason for working hard all of these years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

TE	Transposable Element
polyTE	Polymorphic Transposable Element
1KGP	1000 Genomes Project
LINE	Long Interspersed Element
SINE	Short Interspersed Element
ERV	Endogenous Retrovirus
SNP	Single Nucleotide Polymorphism
AIM	Ancestry Informative Marker
MDS	Multi-Dimensional Scaling
MALD	Mapping by Admixture Linkage Disequilibrium
RMSD	Root-Mean-Square Difference
PBS	Population Branch Statistic

SUMMARY

Transposable elements (TEs) are a class of genes that are characterized by their ability to move (transpose) among locations in the genome. TEs often replicate when they transpose, and over time they can accumulate to very high genomic copy numbers. Accordingly, TEs have had a major impact on the structure, function and evolution of their host genomes. For example, nearly 70% of the human genome sequence is derived from transposable elements (TEs) [1, 2]. This dissertation is focused on the role that recent TE activity has played in shaping genetic variation within and between human populations.

The vast majority of human TE-derived sequences are remnants of relatively ancient insertion events and are no longer capable of transposition. The insertion sites of such inert TE sequences are said to be 'fixed' among human populations, *i.e.*, they are found at the same genomic locations within the genomes of all human individuals. Thus, by definition, fixed TEs do not contribute to human population genetic variation. Nevertheless, until this time, the vast majority of genomic and bioinformatic studies of human TEs have focused on these ancient, fixed TE sequences. This dissertation is distinguished by its focus on human polymorphic TE (polyTE) insertions that vary with respect to their insertion site locations within the genome sequences of different individuals. PolyTEs represent an underappreciated and largely unexplored source of human genetic variation.

There are three main families of TEs that continue to actively transpose in humans, thereby generating insertion site population genetic variation: L1, Alu and SVA [3]. Until very recently, it has not been possible to characterize the genetic variation generated by the activity of these TE families at the scale of whole genomes for multiple individuals within and between populations. For this reason, the impact of TE activity on human evolution has not yet been fully appreciated. This dissertation leverages the convergence of several novel, high-throughput experimental and computational technologies, which together allow for the systematic characterization of genome-wide collections of polyTE insertion genotypes for thousands of individual human genome sequences across scores of distinct population groups. As such, the results reported herein represent the dawn on the population genomics era for human TEs. The research advances detailed in this dissertation are focused on the large-scale characterization of TE polymorphisms and their impact on human evolution.

Research advance 1: Chapter 2 describes an evaluation of the computational techniques that are used to characterize human polyTE insertion sites from whole genome, next-generation sequence data. The corresponding publication represents the first unbiased and comprehensive effort to compare the utility and performance of this class of bioinformatics tools. To do this, we performed a series of controlled benchmarking analyses on 21 recently released computational polyTE detection methods using both validated and simulated human genome sequence data sets. We provide information as to which tools are most reliable along with specific instructions for their installation and use. These

results can help to guide investigators on the optimal use of these programs for human TE research.

Research advance 2: Chapter 3 describes a population genomic analysis of polyTE insertion genotypes as markers of human genetic ancestry and admixture. The corresponding publication represents the first genome-scale evolutionary analysis of polyTE insertion data characterized for thousands of individuals from multiple human populations. To do this, we analyzed 16,192 human polyTE insertions from 2,504 individuals sampled from 26 global populations. We found that polyTE insertion genotypes are reliable population genetic markers that recapitulate known patterns of human evolution. We also demonstrate that polyTE insertion genotypes can be used to make inferences about the patterns of genetic admixture between previously isolated human populations. These results underscore the utility of polyTEs as signals of human genetic ancestry and can help to guide investigators with respect to the selection of specific polyTE insertions that can be used as ancestry informative markers.

Research advance 3: Chapter 4 describes a population genomic analysis of the effects that natural selection has exerted on human polyTE insertions. The corresponding publication represents the first development and application of a genome-wide screen for natural selection on human TE sequences. To do this, we analyzed 14,384 human polyTE insertions from 1,511 individuals sampled from 15 global populations. We developed a novel statistical approach for the detection of selection on TE sequences that combines the

analysis of allele frequencies, phylogenetic inference and time-forward evolutionary simulations. We found that, consistent previous results, the vast majority of human polyTE insertions are constrained by purifying (negative) selection. Nevertheless, we also uncovered a number of cases of polyTE insertions that have increased in population-specific allele frequencies owing to the effects of adaptive (positive) selection. These results illustrate that genetic variation caused by the recent activity of human TEs can provide functional utility for their host genomes.

Research advance 4: Chapter 5 presents a broad prospectus on the implications of genome-scale analyses of human polyTE insertions for population and clinical genetic studies. The corresponding publication represents the first attempt to jointly consider the impact of recent technological developments in genomics, bioinformatics and high-throughput experimental techniques for the study of human TEs. We provide an overview of novel experimental and computation technologies that are used to characterize polyTE insertion sites genome-wide followed by a description of specific areas of potential impact for studies of human evolution and disease. The relevance of recent TE activity to human evolution is considered with respect to both studies of genetic ancestry and natural selection. The impact of recent TE activity on human health is considered for Mendelian disease, common (complex) disease and cancer. These discussions can serve as a guide for future studies of human TEs.

CHAPTER 1. INTRODUCTION

1.1 Transposable elements defined

Transposable elements (TEs) are sequences of DNA that are capable of moving and replicating themselves in the genome. The process of transposition often results in a substantial increase in genomic TE copy number over time. For these reasons, TEs are also sometimes referred to as "jumping genes", "genomic parasites" or "selfish DNA". TEs were first discovered in the maize genome by Barbara McClintock in the 1940s [4]. Initially, TEs did not receive much attention in the field, but they have gradually come to be recognized as major players in genome structure, function and evolution, having been described in virtually all organisms [5]. Due to their replicative nature, TEs can accumulate to occupy a substantial proportion of an organism's genome. For example, nearly 12% of the *Caenorhabditis elegans* [6, 7] genome and over 85% of the *Zea mays* [8] genome are derived from TE sequences. The proliferation of TE sequences in their hosts' genomes can be attributed to an inherent imbalance between their rates of insertion versus removal from the genome [9, 10].

TEs can be broadly divided into two classes (Figure 1) based on their replication mechanisms: 1) DNA transposons and 2) Retrotransposons [11]. The fraction of a host's genome derived from these elements differs from species to species. For instance, the most abundant TEs in *Oryza sativa* genome are DNA transposons, whereas in the human genome, retrotransposons are the most abundant class of elements [12]. DNA transposons are TEs that transpose using a "cut-paste" mechanism; *i.e.*, they excise themselves from the genomic source location using a transposase and then insert into a different target

location via a DNA intermediate. DNA transposons have a characteristic terminal inverted repeat (TIR) sequence at their ends and produce direct sequence repeats at the new site of insertion [12, 13]. DNA transposons are the most common class of TEs found in the genomes of bacteria, where they are referred to as insertions sequences or IS. They are also the most abundant class of TEs found in the genomes of many insects, worms and plants [13].

Retrotransposons transpose via a "copy-paste" mechanism in which the transposition of the source element to the target location occurs via an RNA intermediate. This process of retrotransposition is inherently replicative can lead to a massive genomic accumulation elements. Retrotransposons are further divided into two superfamilies depending on the presence or absence of a long terminal repeat (LTR) sequences at their ends. LTR containing retrotransposons, such as the retrovirus-like elements (RLEs), are evolutionary progenitors of, and similar in structure to, retroviruses [14]. Non-LTR retrotransposons are the second broad group of retroelments, which as the name suggests are characterized by the absence of LTR sequences [15]. These retrotransposons are the most prevalent TE type in mammalian genomes, including the human genome, and include families of Long Interspersed Elements (LINEs) and Short Interspersed Elements (SINEs).

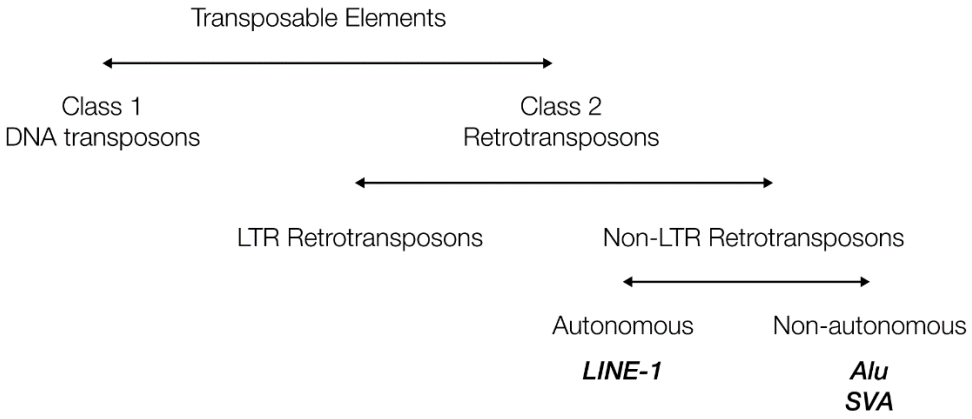

Figure 1 Broad classification of human transposable elements (TEs).

Lastly, TEs can also be divided into autonomous and non-autonomous elements. Autonomous elements are TEs that encode their own transposition machinery (*e.g.*, transposase or reverse transcriptase enzymes), whereas non-autonomous elements depend on enzymes encoded in *trans* by autonomous elements for their transposition. Both DNA trasnposons and retrotransposons include families of autonomous and non-autonomous elements. LINEs are typically autonomous elements, or derivatives thereof, whereas SINEs are obligate non-autonomous elements.

1.2 Transposable elements (TEs) in the human genome

Based on the first draft of the human genome sequence in 2001, it was reported that nearly 45% of the human genome was derived from ancient TE activity [1]. This was believed to be an underestimate as many TE derived sequences were expected to have diverged beyond recognition [1, 14, 16]. A decade later, this estimate of genome TE composition was

revised to 66%-69% accounting for distantly related elements, further underscoring the impact of TEs on the evolution of the human genome [2].

Retrotransposons are the most abundant class of TEs in the human genome, specifically non-LTR retrotransposons (Table 1). The TEs contributing the most mass to the genome are LINEs which altogether make up 20.4% of the genome, with ~868,000 copies. LINE-1 or L1 is the most abundant LINE family in the human genome with an estimated ~516,000 copies, constituting 16.9% of the genome. In terms of numbers of copies, Alus are the most prolific TEs with nearly 1.1 million copies (13.1% of the genome). This is quite remarkable considering that Alus are one of the youngest TE families found in the human genome; Alus arose in primates about 65 million years ago [17] whereas LINEs arose in mammals around 170 million years ago [16]. LTR retrotransposons and DNA transposons are the next most abundant human TE families, making up 8.3% and 2.8% of the human genome, respectively [1].

Table 1 Abundance of different TE types in the human genome.

Type of TE	TE Size	% in Genome	# of Copies
LINE-1	6-8kb	16.89	516,000
Alu	0.3kb	10.60	1,090,000
LTR elements	1.5-11kb	8.29	443,000
DNA transposons	80bp-3kb	2.84	294,000
Others	-	5.78	-
Non-TE	-	55.60	-

TEs are differentially distributed across the human genome sequence. LINE derived sequences are observed at a much higher density in gene-poor, AT-rich regions with a roughly four-fold enrichment. The distribution of LINEs in gene-poor regions is consistent

with their AT-rich target site preference and also fits the expectation that their insertion in intergenic regions is allowed owing to a lower mutational burden. SINEs, on the other hand, show an opposite trend, with a nearly five-fold depletion in AT-rich regions (or enrichment in GC-rich regions). In other words, SINEs are more commonly observed in gene rich regions. This enrichment, however, is skewed towards older Alu families; older families of Alu insertions are enriched in GC-rich regions whereas the newer elements are depleted. Earlier explanations of this observation led to the assumption of some positive selective force acting to preserve Alus in GC-rich DNA [18, 19]. It was later shown to be more likely due to the relative ease with which Alu deletions were tolerated in gene poor AT-rich regions, compared to gene rich GC-regions where Alu deletions via ectopic recombination are observed to be far more deleterious[17, 20-24].

From an epigenetic standpoint, a similar age-based skew is also observed with the enrichment (or depletion) of active and repressive epigenetic marks. Older TE families tend to be enriched with active histone modifications marks compared with younger TE families [25]. This observation was reported to be more in line with the 'exaptation hypothesis' [26] which argues for the repurposing of older elements to provide regulatory sequences for their host genomes.

1.3 Active human transposable elements (TEs)

Over 99% of the TE derived sequences in the human genome are ancient remnants of past insertional activity that are no longer capable of transposition. These functionally inert TE sequences are present at fixed locations in the human genome that do not vary between

individuals. In other words, these relatively ancient TE derived sequences do not contribute to insertion site genetic variation within or between human populations. However, there are three main families of TEs that are still actively transposing in the human genome [3] – L1 [27, 28], Alu [29, 30] and SVA [31, 32]. L1, or LINE-1, are autonomous, non-LTR retrotransposons. A full-length L1 is nearly 6kb long and contains an internal RNA polymerase II promoter, a 5' UTR, two ORFs, a 3' UTR and a polyA tail (Figure 2). The first ORF codes for an RNA-binding protein, and the second ORF codes for endonuclease and reverse transcriptase enzymes. L1 elements rely on a host genome encoded RNA polymerase to create an RNA copy of themselves and then use their own enzymes to reverse transcribe the mRNA and integrate the resulting DNA copy in the genome using the mechanism known as target-primed reverse transcriptase (TPRT) [33-35]. Alus and SVAs belong to the non-autonomous, non-LTR retrotransposons are transposed in *trans* via L1 encoded transposition machinery [36, 37]. Alus are 7SL RNA derived SINEs that are typically 300bp in length (Figure 2) [38, 39]. SVAs are composite elements made up of SINE (Alu-like), VNTR (Variable Number Tandem Repeat) [40, 41] and HERV-K10-like elements and can vary from 100-2,000bp in length (Figure 2) [42].

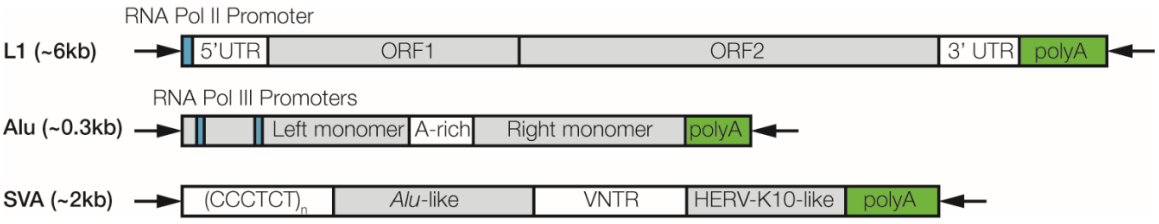

Figure 2. Genomic structures of the three main active TE families in the human genome (not to scale).

Germline transposition of active human TEs can lead to human genetic variation in the form of TE insertion presence/absence patterns between individuals. Germline transposition of these elements is relatively rare. According to recent estimates in humans, *de novo* Alu, L1, and SVA germline insertions occur once for every 21, 212, and 916 live births, respectively [43]. Such transposition events in the human genome can result in severe phenotypic consequences [44, 45]. However, not all TE insertions are deleterious in nature, as polymorphic TE (polyTE) insertions have been reported in a number healthy human individuals, create insertion alleles that segregate among human populations, and can accumulate to high allele frequencies [3, 17, 37, 46-56].

1.4 Computational detection of polymorphic transposable element (polyTE) insertions

With the advent of next-generation sequencing (NGS) technologies, it has become possible to characterize genome-wide patterns of polyTE insertions using computational analysis of whole genome sequence data [57]. A number of methods for the computational detection of TE insertions from NGS data have been released in the last few years. All of these bioinformatic methods operate on essentially the same paradigm of analyzing short, paired-end sequence reads mapped to the reference genome sequence. TE insertion detection methods focus on two types of read mapping relationships - discordant read pairs (DPs) and split (or clipped) read pairs (SRs) (Figure 3). DPs are read pairs where one member of the pair maps uniquely to the reference genome sequence, while the other member of the pair maps ambiguously to a TE family. SRs consists of reads that map to the junction

of the reference genome sequence and the inserted non-reference TE sequence. In other words, the reads are split (clipped) with part of the read mapped to a unique sequence and the other part mapped to a repetitive TE sequence. DPs and SRs will show distinct mapping characteristics when mapped to a reference genome that lacks the polyTE insertion. These distinct mapping characteristics are used by the polyTE detection programs, together with the partial mapping to active TE sequences, to identify and locate polyTE insertions [58].

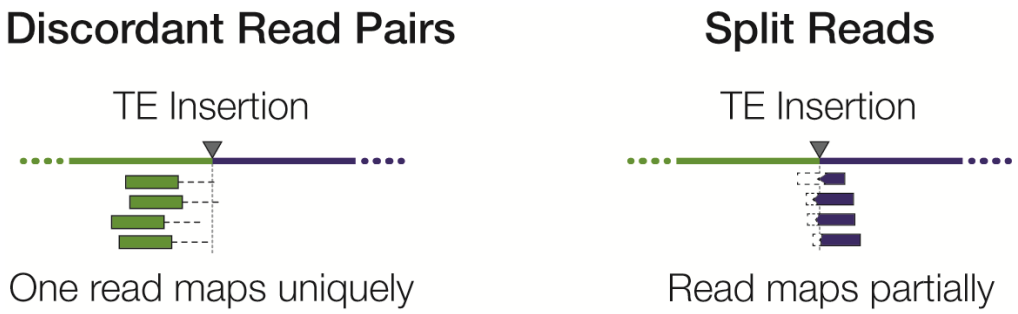

Figure 3 Read mapping types frequently analyzed for computational TE detection from whole genome sequencing data.

This algorithmic paradigm for computationally detecting TEs was first introduced in 2011 [59], and since then, it has been adopted by a number of developers and applied on a diverse set of organisms including rice [60], fruit fly [61] and mouse [62]. Despite these efforts, the development of computational tools for the detection of non-reference TE insertions from NGS data remains in its infancy, and the performance of these tools has only been compared by the tool's respective developers. Their actual performance has never been validated using a known set of genome wide polyTEs, which raises serious concerns with their accuracy and reliability. This also presents a significant challenge for the TE research community when it comes to selecting the best tool(s) for computationally predicting non-

reference TE insertions from NGS data. An independent and systematic attempt to compare the performance of polyTE detection tools is much needed and will immensely benefit the TE research community.

1.5 Evolutionary implications of human polyTE insertions

1.5.1 Ancestry informative markers

As described in section 1.3, ongoing transposition activity of Alu, L1 and SVA elements in the human germline yields TE insertion sites that are polymorphic among human populations. Such polyTE insertion sites have a number of features that make them potentially valuable sources of ancestry informative markers (AIMs) that can be analyzed to make inferences about the evolution of their host genomes. Given the size of TE insertions compared to that of the genome, coupled with the low overall retrotransposition rate, the probability of observing independent insertions of different TEs at the same chromosomal location is diminishingly low. Furthermore, once inserted, TEs rarely undergo complete deletion or rearrangement [47]. These two features together make polyTE insertions stable genetic markers that are free of homoplasies – *i.e.*, identity of state that is not due to common descent, which is far more commonly seen for single nucleotide polymorphism (SNP) markers. Finally, the ancestral state for any polyTE insertion can assumed to be the absence of an insertion, an additional feature that provides added utility for evolutionary inference. PolyTEs, especially Alus, are also practical genetic markers as they can be rapidly and accurately typed via PCR-based assays. Overall, TE insertion site polymorphisms present an intriguing and unappreciated aspect of human genetic variation,

with an impact on human ancestry and evolution that remains to be evaluated at the genome-wide scale.

The quest for ancestry informative polyTE markers has been greatly limited due to the lack of a large, population-scale, genome-wide dataset of characterized polyTE insertion sites. Researchers in the past have relied on either literature or database surveys for the selection of potential polyTE insertion sites to serve as ancestry markers [3, 17, 37, 46-56]. Despite the rather *ad hoc* nature of previous polyTE marker selection approaches, these studies have been successful in demonstrating the ability of polyTEs to serve as AIMs. But it has not yet been possible to evaluate the relationship between TE polymorphism and human evolution in a systematic and unbiased way. A comprehensive genome-wide survey of polyTE insertion loci and their utility as AIMs also remains to be tested.

1.5.2 Selection

There is abundant evidence suggesting that insertions of TEs in the human genome are highly deleterious. In fact, the identity of repetitive L1 sequences as a family of TEs was first discovered via analysis of a hemophilia A patient with a deleterious, novel L1 insertion in the *F8* (Coagulation Factor VIII) gene [28]. Several subsequent studies have implicated TE insertions in a number of other genetic diseases such as hemophilia B [63], cystic fibrosis [64], Apert syndrome [65], X-linked agammaglobulinaemia [66] and in a variety of different cancer types including testicular cancer [67], germ cell tumors [68] and breast cancer [69, 70]. The numerous studies reporting deleterious effects of TE insertions can be considered to be consistent with the selfish DNA theory, with respect to the notion that

TEs are genomic parasites that impose a mutational burden on their host, and also point to a role for purifying selection in countering their unchecked spread.

However, there is also abundant evidence that TE sequences can serve as a creative force in evolution by providing a substrate for the emergence of novel functions. In the years since the publication of the draft human genome sequence, there have been many studies that have demonstrated how formerly selfish human TE sequences have been exapted [71], or domesticated [72], to play a functional role in the human genome. This has been seen most often in the context of regulatory sequences [73]. Human TE sequences have been shown to provide a wide variety of gene regulatory sequences including promoters [74-76], enhancers [77-81], transcription terminators [82] and several classes of small RNAs [83-85]. Human TE sequences can also affect host gene regulation via changes in the local chromatin environment [19, 86-90].

It is important to note that all of the aforementioned studies report results for ancient human TE sequences that are no longer capable of transposition and are thereby located at fixed genomic locations. At this time there is far less evidence that active human TE families can also be positively selected based on some functional utility for the host genome. There is abundant evidence of adaptive evolution of polyTEs in *Drosophila* [91-95] along with studies that show the regulatory potential of polyTEs in mice [96]. Moreover, at this time there is tentative evidence to suggest that human polyTEs have been subject to positive (adaptive) selection [97], but this has not been tested in a systematic and genome-wide manner. With the recent developments in genomics and bioinformatics, it is now beginning to be possible to unambiguously evaluate the effects of natural selection (both negative and positive) on human polyTEs.

1.6 Population genomics of human polyTEs

Until recently, obtaining a genome-wide set of polyTE insertion sites was a daunting task. Consequently, most researchers focused on small sets of polyTE insertion sites identified by literature or database searches. In November 2014, the 1000 Genomes Project (1KGP) released the first genome-wide set of human polyTE presence/absence insertion genotypes characterized for 2,504 individuals from 26 different populations as part of their phase III variant release (Figure 4) [42].

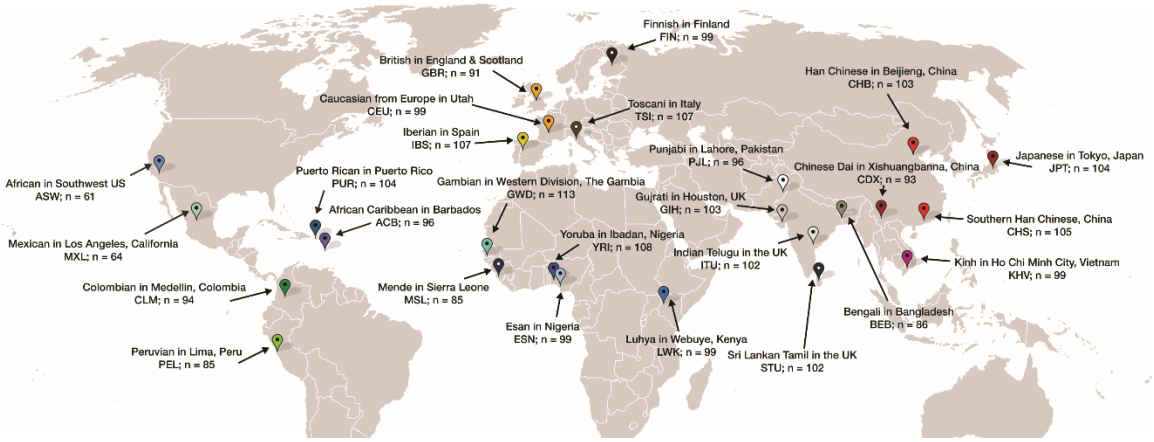

Figure 4 Human populations characterized in the phase III release of the 1000 Genomes Project.

The 1KGP utilized the recently developed computational tool MELT to predict the genome-wide polyTE insertion site genotype calls from 2,504 individuals surveyed as part of the project. The accuracy of the predictions of the TE calls was measured by random PCR validations and Sanger sequencing [42, 98]. Additional verification was also performed using the genome sequence of a HapMap/1KGP CEU (European) female sample 'NA12878' that was deeply characterized using long read sequencing technology

with the Pacific Biosciences Single Molecule Real Time method. PolyTE insertions obtained from this *de novo* assembled genome were considered as experimentally validated since the long reads spanned polyTE insertion sites [42]. MELT was found to have a 98% genotype concordance at with a 96% sensitivity and 4% site FDR [42]. A total of 16,192 polyTE insertion site genotypes characterized by the MELT method were reported across all sampled 1KGP individuals. An average human genome was shown to contain approximately 4-5 million SNPs and ~1,100 polyTE insertions [99]. In terms of number of polymorphic bases in the human genome, this accounts to 4-5 million bases from SNPs and nearly 1.1 million bases from polyTE insertions, representing a substantial amount of human diversity that has never been studied before. The lowest number of polyTE insertions observed in any individual was seen for a sample donor from PJL (Pakistan) with 543 insertions, half of the number of insertions relative to the global average (Figure 5A). The individual with the most polyTE insertions belonged to the admixed American group ASW (African Americans from South West US) with 1,715 insertions, 50% more than the global average (Figure 5B). The genome-wide collection of polyTE insertion site genotypes is enabling TE researchers, including our own group as reported in this thesis, to analyze a previously unexplored dimension of human genetic diversity and to answer a number of unanswered questions in the fields of human TEs, genomics and evolution.

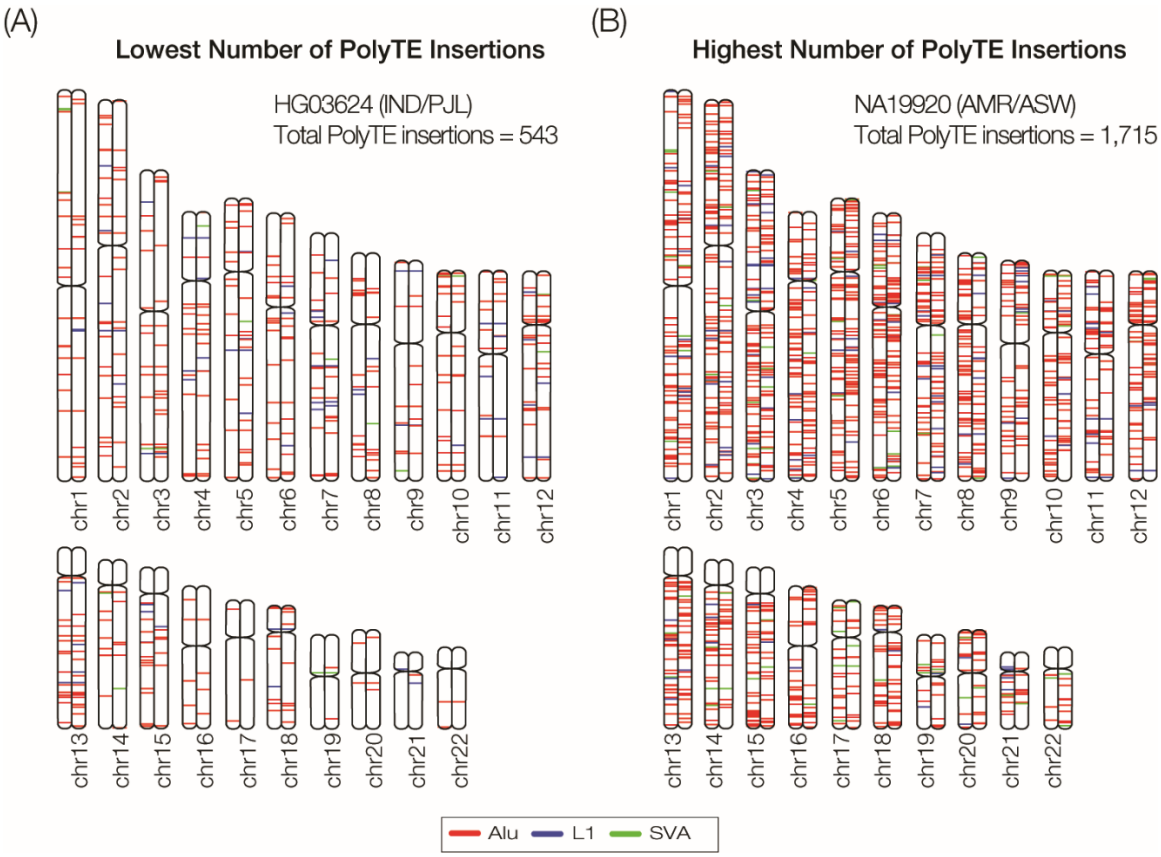

Figure 5 The lowest and highest number of polyTE insertions seen for the 1000 Genomes populations.

Chromosomal locations of polyTE insertion sites are shown for the individuals with the (C) lowest and (D) highest numbers of insertions.

CHAPTER 2. BENCHMARKING COMPUTATIONAL TOOLS FOR POLYMORPHIC TRANSPOSABLE ELEMENT DETECTION

2.1 Abstract

Transposable elements (TEs) are an important source of human genetic variation with demonstrable effects on phenotype. Recently, a number of computational methods for the detection of polymorphic TE (polyTE) insertion sites from next-generation sequence data have been developed. The use of such tools will become increasingly important as the pace of human genome sequencing accelerates. For this report, we performed a comparative benchmarking and validation analysis of polyTE detection tools in an effort to inform their selection and use by the TE research community. We analyzed a core set of seven tools with respect to: ease of use and accessibility, polyTE detection performance, and runtime parameters. An experimentally validated set of 893 human polyTE insertions was used for this purpose, along with a series of simulated data sets that allowed us to assess the impact of sequence coverage on tool performance. The recently developed tool MELT showed the best overall performance followed by Mobster and then RetroSeq. PolyTE detection tools can best detect Alu insertion events in the human genome with reduced reliability for L1 insertions and substantially lowered performance for SVA insertions. We also show evidence that different polyTE detection tools are complementary with respect to their ability to detect a complete set of insertion events. Accordingly, a combined approach, coupled with manual inspection of individual results, may yield the best overall performance. In addition to the benchmarking results, we also provide notes on tool installation and usage as well as suggestions for future polyTE detection algorithm development.

2.2 Introduction

2.2.1 Polymorphic TEs in the human genome

Transposable elements (TEs) are mobile DNA sequences that are capable of accumulating to high copy numbers in their host genomes. Indeed, it has been estimated that ~50-70% of the human genome is made up of TE-derived sequences [1, 2]. These TE-derived sequences represent scores of families that have accumulated copies in the genome over many millions of years, a very small fraction of which remain transpositionally active [3]. The main active families of human TEs are L1 [27, 28], Alu [29] and SVA [31, 32]. All three of these families correspond to retrotransposons that transpose via reverse transcription of an RNA intermediate. L1 elements are a family of long interspersed nuclear elements (LINEs) [33, 34], which are considered to be autonomous in the sense that they encode the enzymatic machinery necessary to catalyze their own retrotransposition [35]. Alu and SVA elements are non-autonomous TEs, which are transposed in *trans* by the L1 machinery [36, 37]. Alu elements are so-called short interspersed nuclear elements (SINEs) that evolved from 7SL RNA [38, 39], and SVAs are composite elements that are made up of human endogenous retrovirus sequence, simple sequence repeats and Alu sequence [40, 41].

Transpositional activity of active human TE families is an important source of genetic variation that can have severe phenotypic consequences. Mutations cause by TE insertions are known to cause a number of genetic diseases, including several kinds of cancer [45, 100, 101]. Alu insertions are linked to breast cancer and cystic fibrosis; L1 insertions can cause colon cancer and haemophilia A, and SVA insertions lead to leukemia and X-linked

dystonia-parkinsonism. Active human TEs are also relevant to population genomic studies since polymorphic TE (polyTE) loci can serve as valuable genetic markers for studies of human ancestry [42, 102]. Given the relevance of TE activity to human clinical and population genomics, the ability to systematically characterize polyTEs from accumulating human genome sequences will be critical.

Over the last several years, a number of computational tools have been developed for the characterization of polyTE insertions based on the analysis of next-generation sequence data [58]. Computational polyTE detection tools will become increasingly important for studies of human genome sequence variation owing to the emergence of numerous efforts to characterize thousands of whole genome sequences. The 1000 Genomes (1KG) Project was the first effort of this kind [99, 103], and the recent Phase III data release contains a complete catalog of >16,000 polyTE loci among 2,504 individuals [42]. The National Heart, Lung and Blood Institute of the US National Institutes of Health has an initiative underway to sequence whole genomes for 70,000 individuals [104], and the Sanger Institute in the United Kingdom is sequencing 100,000 human genomes [105]. These are just a few of many such initiatives that are underway around the world.

Despite the accumulation of data from these massive sequencing efforts, the development of computational tools for the detection of polyTE insertions from next-generation sequence data remains in its infancy, and there has yet to be a systematic attempt to compare the utility and performance of polyTE detection tools. In this report, we present a comparative benchmarking and validation analysis of computational tools for polyTE detection. We have focused this analysis on human genome sequences owing to their clinical importance and impending abundance. In addition, the presence of an

experimentally validated set of polyTE loci for a single human individual provides a valuable resource for tool benchmarking and validation. This study represents a practical evaluation of polyTE detection tools, with an eye towards both users and developers, rather than a comprehensive review of TE sequence analysis tools, which have been covered in depth elsewhere [3, 57, 58].

2.2.2 Polymorphic TE detection tools

The benchmarking study reported here concerns only polyTE detection tools, rather than TE discovery and annotation tools [57] or general structural variant detection tools [106]. TE annotation tools, such as RepeatMasker [107] or CENSOR [108], typically rely on the comparison of TE consensus sequences to assembled genome sequences to characterize the genomic locations, and (sub)family identities, of TE-derived sequences. The vast majority of TE-derived sequences in the human genome are the remnants of ancient insertion events, which are no longer capable of transposition and reside at fixed locations that do not differ between individual genomes. More recent transpositional activity of polyTE families generates insertions that differ between individuals. Detection of such polyTE loci requires different kinds of computational tools, which utilize (re)sequencing data by analyzing the locations to which sequence reads map to a genome reference sequence. This class of computational tools has only been recently developed and has yet to be systematically compared and benchmarked.

We chose a total of seven polyTE detection tools for comparative benchmarking and validation (Table 2). We chose these tools based on a number of criteria by which we

attempted to pre-assess their viability and potential for use by the TE research community: 1) tools that are both recently released (2013 or later) and currently maintained, 2) tools that have been evaluated using actual or simulated human genome next-generation sequence data, 3) tools that were used for polyTE detection in the 1KG Project, and 4) tools developed for other model organisms and have been directly compared with human polyTE detection tools. The seven tools that fit these criteria are listed in Table 2 along with a qualitative assessment of their relative ease of installation, ease of use and the comprehensiveness of their manuals. We provide extended usage details on each of these tools in the Appendix A including exact commands with parameters and input files used. We also provide notes with respect to what is needed to install and run each program (*e.g.,* dependencies) along with brief descriptions of any issues we encountered with their use. Finally, we note cases where use of the tools entailed direct communication with their developers, and the adjustments that were made to facilitate their execution.

Table 2 List of polyTE detection tools benchmarked in this study.

The tools are compared with respect to their ease of installation, ease of use and the comprehensiveness of their manual.

Tool	Ref	Year	Algorithm[1]	Reported Testing Set[2]	Website
MELT	Unpublished		DP/SR	1KG	http://melt.igs.umaryland.edu/
ITIS	[109]	2015	DP/SR	*M. truncatula*	https://github.com/Chuan-Jiang/ITIS
TEMP	[110]	2014	DP	1KG; Simulated	https://github.com/JialiUMassWengLab/TEMP
Mobster	[111]	2014	DP/SR	1KG; EGA	http://sourceforge.net/projects/mobster/
Tangram	[112]	2014	DP/SR	1KG; Simulated	https://github.com/jiantao/Tangram
RetroSeq	[62]	2013	DP	1KG	https://github.com/tk2/RetroSeq
T-lex2	[61]	2014	RM/RD	1KG; *DM*	http://petrov.stanford.edu/cgi-bin/Tlex.html

Tool	Lang[3]	Ease of Installation[4]	Ease of Use	Manual	Open Source	VCF Output[5]
MELT	Java	Easy	Easy	Detailed	No	Yes
ITIS	Perl	Easy	Easy	Detailed	Yes	No
TEMP	Perl	Moderate	Moderate	Detailed	Yes	No
Mobster	Java	Easy	Easy	Detailed	No	No
Tangram	C++	Moderate	Difficult	Brief	Yes	Yes
RetroSeq	Perl	Easy	Easy	Detailed	Yes	Yes
T-lex2	Perl	Difficult	Difficult	Moderate	Yes	No

[1]Algorithmic paradigm used by the tool: DP=Discordant read pairs, SR=Split/Clipped reads, RM=Read Mapping, RD=Read Depth

[2]Test data set used for previously reported validation: 1KG=1000 Genomes Project, EGA=European Genome-phenome Archive, DM=D. melanogaster

[3]Coding language used for the tool development

[4]Includes installation of the program and all required dependencies

[5]Whether or not the program produces a variant call format (VCF) output

While the considerations we used to pre-select polyTE detection tools for analysis here may be somewhat subjective, we feel that the collection of tools benchmarked for this report represents the current state-of-the art for polyTE detection. Readers should be aware that more exhaustive lists of polyTE detection tools have been reported in a recent review

of such detection methods [58] as well as an older review that covered a broader range of TE sequence analysis tools [57]. In addition, online lists of polyTE detection tools can be found on the TE tools @ Bergman Lab website [113] and on the OMICtools website [114]. While the lists found in these papers and websites are far more inclusive than the set of tools we analyze here, they do not provide any indication of utility or performance for the tools or any practical guide for tool selection and use. Here, we have opted for a deeper analysis of a core set of tools, which we hope can serve as a reliable guide for investigators who interested in TE discovery as well as those who may be inclined to pursue further algorithm development in this area.

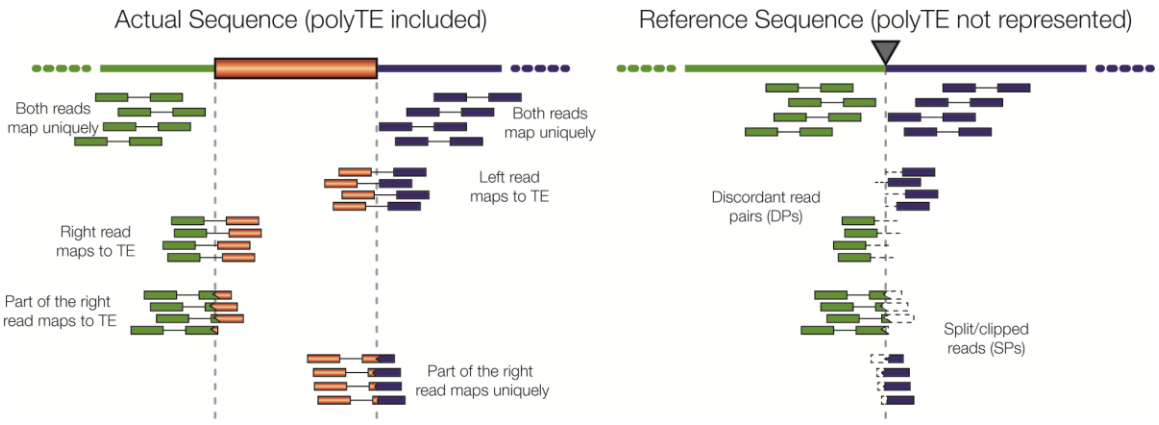

Figure 6 Detection of polyTE insertions using next-generation sequence data.

Two schemes are shown illustrating how paired-end read mapping information is used for the detection of polyTE insertion sites. One scheme shows the actual sequence being characterized with the polyTE insertion present and the other scheme shows the reference sequence that lacks the polyTE insertion. Sequence reads generated from the actual sequence will be mapped to the reference sequence as shown. There are three classes of locally mapped reads that inform polyTE detection: 1) both reads in a pair map uniquely, 2) discordant pairs (DP) where one read maps uniquely and one read maps to a repetitive TE sequence, and 3) split/clipped pairs (SP) where part of one read maps uniquely and the other part maps to a repetitive TE sequence. The presence of DPs and SRs, along with the mapping distance between their paired reads, is used in the prediction of polyTE insertion sites.

All of the polyTE detection tools analyzed here operate on the same basic algorithmic paradigm for the analysis of short, paired-end sequence reads mapped to a reference genome sequence (Figure 6). There are two particularly important classes of reads that point to the presence of a polyTE insertion relative to a reference genome sequence that lacks an insertion at that locus. These are so-called discordant read pairs (DPs) and split, or clipped, read pairs (SRs) (left panel of Figure 6). DPs are read pairs where one member of the pair maps uniquely to the reference genome sequence, while the other member of the pair maps ambiguously to members of an active TE family. SRs contain one read that maps to the junction of the reference genome sequence and the inserted polyTE sequence. In other words, the reads are split (clipped) with part of the read mapped to unique sequence and the other part mapped to a repetitive TE sequence. DPs and SRs will show distinct mapping characteristics when mapped to a reference genome that lacks the polyTE insertion (right panel Figure 6). These mapping characteristics are used by the polyTE detection programs, together with the partial mapping to active TE sequences, to identify and locate polyTE insertions. Various programs also incorporate additional sources of information, *e.g.*, read depth and prior information about polyTE insertion locations, but the DP and SR paradigm is the essence of these algorithms.

2.3 Materials and Methods

2.3.1 Benchmarking sequence data sets

As previously discussed, we focused our polyTE detection tool benchmarking and validation efforts on human genome sequences. To do this, we evaluated a series of actual and simulated next-generation human genome sequence data sets. The actual next-generation sequence data that we analyzed was characterized from the HapMap/1000 Genomes CEU (European) female sample 'NA12878'. This individual corresponds to the sample that has been analyzed extensively as part of the Genome in a Bottle Consortium project, which aims to validate tools for human genome sequence variant calling [115]. As such, it represents the most reliably characterized individual human genome sequence in existence. This sample was sequenced to low coverage (5.7x) in Phase I, and high coverage (95.6x) in Phase II, of the 1KG Project using Illumina short read sequencing technology [103]. Sequence reads for these two runs were mapped to the human genome reference sequence as previously described [103], and the read-to-genome alignment files (*i.e.*, the BAM files) were obtained from the 1KG website [116] for use with the polyTE detection tools evaluated here.

Most importantly, with respect to the validation of polyTE detection tools, this same sample was also characterized using the Pacific Biosciences (PacBio) long read sequencing technology by members of the 1KG Project Structural Variation Group [42]. The use of PacBio sequencing technology allowed this group to unambiguously characterize the insertion sites for 893 human polyTEs in the NA12878 genome sequence, since the long reads span (or can be readily assembled across) polyTE insertion sites. The resulting

validated polyTE insertion sites (generously provided by Dr. Ali Bashir, Icahn School of Medicine at Mount Sinai) were used to assess the performance of the polyTE detection tools benchmarked here.

We also benchmarked the polyTE detection tools using simulated human genome sequence data in an effort to more thoroughly explore the effect of different sequence coverage levels on the performance of the tools. The simulated data was generated by randomly inserting AluY, L1 and SVA consensus sequences, taken from RepBase 14.02 [117], into the autosomes of the human genome reference sequence (build 19, GRCh37) using a custom written Perl script. Each simulated polyTE insertion had a chance of undergoing stochastic single base mutations at a rate of up to 15% in an effort to reflect naturally occurring variation among dispersed TE copies. Simulated insertions included poly-A tails and target site duplications, as these features are used by some polyTE detection tools. A total of 893 polyTE insertions were created with the proportions of AluY, L1 and SVA following the reported worldwide genomic averages of 915, 128 and 51 insertions respectively [42, 102]. Having created an *in silico* set of polyTE insertions in this way, paired-end reads were then simulated using the ART simulator [118] with the Illumina MiSeq profile, read length of 150bp, mean fragment length of 500 bp and a standard deviation of 10 bp. Read simulation was done across a range of approximate coverage values: 5x, 10x, 15x, 30x and 50x (Table 3). Simulated reads were mapped to the human genome reference sequence (build 19, GRCh37) using the program BWA [119], and all subsequent file format conversions and sorting were done using SAMtools [120].

Table 3 Actual and simulated data sets used for benchmarking polyTE detection tools.

Data set[1]	Source	# of Reads[2]	Bases Sequenced	Read Length	Coverage[3]
NA12878 Low	1KG Phase I	172,724,240	17,445,148,240	101	5.74x
NA12878 High	1KG Phase II	2,873,647,625	290,238,410,125	101	95.59x
Sim5x	Simulated Data set	89,510,496	13,426,574,400	150	4.66x
Sim10x		179,023,214	26,853,482,100	150	9.32x
Sim15x		268,528,918	40,279,337,700	150	13.98x
Sim30x		537,056,924	80,558,538,600	150	27.96x
Sim50x		895,112,564	134,266,884,600	150	46.60x

[1] Actual and simulated data sets used for benchmarking (as described in the text)
[2] Total number of sequence reads present in each data set
[3] Genomic coverage (*i.e.*, sequencing depth) for each data set

2.3.2 Benchmarking and validation parameters

The seven polyTE detection tools shown in Table 2 were run using the low and high coverage actual human genome sequence read data sets from the NA12878 sample as well as the five simulated read data sets across a range of coverages (Table 3). The tools were run on a high performance server with 512GB of RAM and 4 10-core Intel Xeon 2.8GHz processors. The details for how each tool was run, along with notes guiding their installation and use, are shown in the Appendix A. The tools were benchmarked and validated according to two broad performance categories: 1) polyTE detection performance and 2) runtime parameters. The details of the results of this comparative analysis are shown in Table 4.

Table 4 Benchmarking and validation results for seven polyTE detection tools.

The tools were evaluated broadly for polyTE detection performance and runtime parameters as described in the text.

Data	Tool	PolyTE Detection Performance							Runtime Parameters			
		Total Predictions [1]	Correct Prediction: Exact [2]	Correct Prediction: ≤100 bp [3]	Correct Prediction: ≤1 Kb [4]	TP [5]	FP [6]	FN [7]	CPU time [8]	Wall time [9]	Peak RAM [10]	% CPU [11]
NA12878 Low	MELT	1,189	853	862	862	862	327	31	13.5	18.6	19.07	111
	Mobster	1,035	39	651	678	651	384	242	18.3	65.2	101.43	76
	RetroSeq	749	5	408	515	408	341	485	96.6	83	0.39	121
	TEMP	4,928	0	31	45	31	4,897	862	36	42.9	2.19	97
	Tangram	3,186	172	411	413	411	2,775	482	384.8	123.8	98.3	322
	ITIS	237	37	77	184	77	160	816	2,316.20	689.9	28.67	347
	T-lex2	*Process Killed. Reason: Process not finished within a week*							> 1 week	-	-	-
NA12878 High	MELT	179	45	47	47	47	132	846	232.3	360.6	80.12	92
	Mobster	1,572	303	819	825	819	753	74	449.5	426	118.55	156
	RetroSeq	4,404	21	850	859	850	3,554	43	1,889.40	1,653.5	1.67	124
	TEMP	1,109	2	49	87	49	1,060	844	948.9	1,187.1	150.14	92
	Tangram	*Process Killed. Reason: Exited with error message. Reported problem.*							6,611.20	3,128.9	261.61	221
Reported	MELT	990	807	807	807	807	183	86	-	-	-	-
	Mobster	1,250	352	800	805	800	450	93	-	-	-	-
	RetroSeq	1,252	18	791	799	791	461	102	-	-	-	-
	Tangram	1,553	250	828	837	828	725	65	-	-	-	-
Sim5x	MELT	304	22	264	294	264	40	628	6.1	16.8	13.98	95
	Mobster	322	4	271	300	271	51	621	11.31	14.3	10.42	120
	RetroSeq	662	3	348	631	348	314	544	42.45	35.32	0.39	124
	ITIS	66	0	23	62	23	43	870	2,621.10	1,057.9	57.14	261
	TEMP	*No Predictions*							2.37	2.47	2.04	98
	Tangram	*Process Killed. Reason: Exited with error message. Reported Issue.*							180.4	71.58	51.02	256
Sim10x	MELT	416	35	396	402	396	20	496	11.45	12.85	14.92	122
	Mobster	505	7	406	439	406	99	486	17.89	18.71	10.43	110
	RetroSeq	769	5	434	730	434	335	458	96.05	78.35	0.39	126
	ITIS	172	0	35	160	35	137	857	4,247.16	1,248.6	57.14	352
	TEMP	*No Predictions*							5.06	5.26	2.04	99
	Tangram	*Process Killed. Reason: Exited with error message. Reported Issue.*							343.34	143.78	102.04	246
Sim15x	MELT	484	51	460	467	460	24	432	16.84	20.72	12.97	118
	Mobster	570	9	460	493	460	110	432	26.36	39.33	10.53	124
	RetroSeq	734	11	489	734	489	245	403	113.5	92.08	0.39	126
	ITIS	256	0	42	241	42	214	850	6,985.06	1,937.7	57.14	372
	TEMP	*No Predictions*							6.54	6.67	2.04	100
Sim30x	MELT	542	67	509	520	509	33	383	28.55	27.75	12.05	112
	Mobster	439	16	405	413	405	34	487	44.89	35.03	10.52	134
	RetroSeq	804	14	507	738	507	297	385	260.82	216.38	0.39	123
	ITIS	399	0	49	352	49	350	843	13,286.9	3,185.7	57.14	428
	TEMP	*No Predictions*							12.19	12.45	2.04	100
Sim50x	MELT	562	68	527	539	527	35	365	45.42	52.14	41.66	102
	Mobster	593	14	505	515	505	88	387	60.78	69.95	34.05	107
	RetroSeq	828	19	489	742	489	339	403	398.71	323.13	0.39	126

For polyTE detection performance, the locations of predicted polyTE insertion for each tool were compared to known insertion sites from the actual and simulated data sets. PolyTE insertion site locations that were predicted within 100bp of a known insertion site were counted as true positives (*TP*). Predictions that fell outside this range were counted as false positives (*FP*), and known polyTE insertion sites that did not have any prediction within 100bp were counted as false negatives (*FN*). The resulting *TP*, *FP* and *FN* counts were used to compute Precision, Recall and F1-Scores, as metrics of the relative performance of the polyTE detection tools. Precision (also known as positive predictive value) is computed as *TP/(TP+FP)*, and it characterizes the ability of the tool to reject false insertion predictions. Recall (also known as sensitivity or true positive rate) is computed as *TP/(TP+FN)*, and it characterizes the ability of the tool to predict true insertions. Finally, the F1-Score (also known as the F-measure) is computed as the harmonic mean of Precision and Recall, *2x[(PrecisionxRecall)/(Precision+Recall)]*, and it is used here to measure the overall polyTE detection performance of each tool.

Runtime parameters measure the amount of time and computational resources used by the polyTE detection tools. The CPU time is the amount of processor time used by the tool, whereas the Wall time is the actual wall clock time that the tool takes to finish. Peak RAM is the maximum amount of memory occupied by the tool over the course of its run, and the %CPU is the percentage of the cores that the tool was able to utilize.

2.4 Results and Discussions

2.4.1 PolyTE detection performance

The relative performance of the seven polyTE detection tools evaluated here is shown in terms of Precision, Recall and the F1-Scores for the actual low and high coverage human genome sequence data sets analyzed here (Figure 7). The unpublished tool MELT shows the best overall performance on the low coverage (5.7x) data set followed by Mobster and then Retroseq. Tangram shows intermediate performance and then there is a precipitous drop off to the next set of three tools, all of which show poor or no performance. Results for the program T-lex2 are not shown here as it took over four weeks to run and predicted over 300k insertions. The superior performance of MELT (97% of all polyTE insertions detected) on this data set is consistent with the fact that it was the program used for the 1KG Project from which the validation data were derived [42], and the tool incorporates prior information in the form of known human polyTE insertion sites. The empirical performance of MELT measured via the current benchmarking analysis is similar to what has been previously reported as opposed to the other tools evaluated here, which tend to show previously reported performance levels that are substantially higher than those observed here.

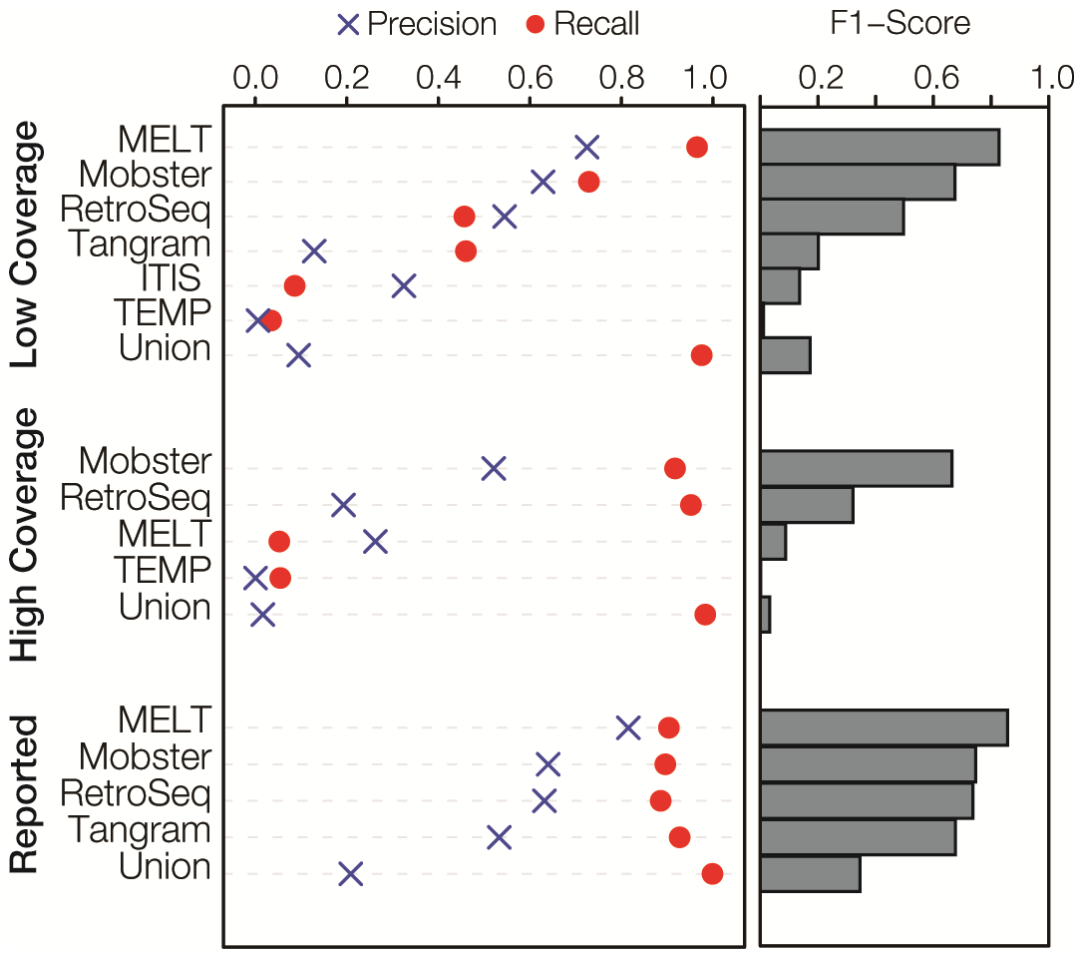

Figure 7 Overall polyTE detection tool performance.

Precision, Recall and F1-Scores are shown for the different polyTE detection tools evaluated here for the low (5.7x) and high (95.6x) coverage human genome sequence NA12878. The same parameter values are shown based on previous reports on these tools. The union of predictions made by all tools under consideration is shown for each category. For each category, the tools are ranked according to the F1-Score, which provides an overall measure of performance.

Surprisingly, the results of the polyTE detection tool evaluation on the high coverage (95.6x) data set indicate that additional sequence coverage can be positively misleading with respect to performance. More data in this case is not better. Only four of the seven tools evaluated here were able to successfully run on the high coverage data set. In

addition, all of the tools showed worse performance on the high coverage data set compared to the low coverage data set. This is based on the fact that all of the tools, except for MELT, predicted substantially higher numbers of polyTE insertions and accordingly had higher numbers of false positives (Table 4). MELT gave an error for Alu and SVA predictions for the high coverage data set and was only able to predict the longer L1 insertions for these data. Mobster and Retroseq show the best performance for this data set, but both of these tools also have high numbers of false positives and accordingly low Precision. In summary, none of these tools work reliably for such a high coverage data set, and users should be cautioned against applying them to such data sets. This problem is mitigated by the fact that it is highly unlikely, at least at this time, that many whole human genome sequences will be sequenced to this depth. Nevertheless, these results underscore the fact that polyTE prediction still remains an inexact science.

We also compared the performance of the polyTE detection tools for Alu, L1 and SVA separately on the low and high coverage data sets (Figure 8). Overall, the three most reliable tools (MELT, Mobster and RetroSeq) work best on Alu elements, followed by L1 and then SVA, which shows the poorest performance by far. Alu insertions are detected with relatively high Precision and Recall in the low coverage data set; L1 insertions have relatively high Recall but much lower Precision, whereas SVAs are low for both Precision and Recall. MELT showed the most uniformly strong performance across all three polyTE families. Alu elements are also distinguished by the fact that the vast majority of insertions can be found by all three of the best methods, whereas there is no single SVA insertion that is found by all of these methods.

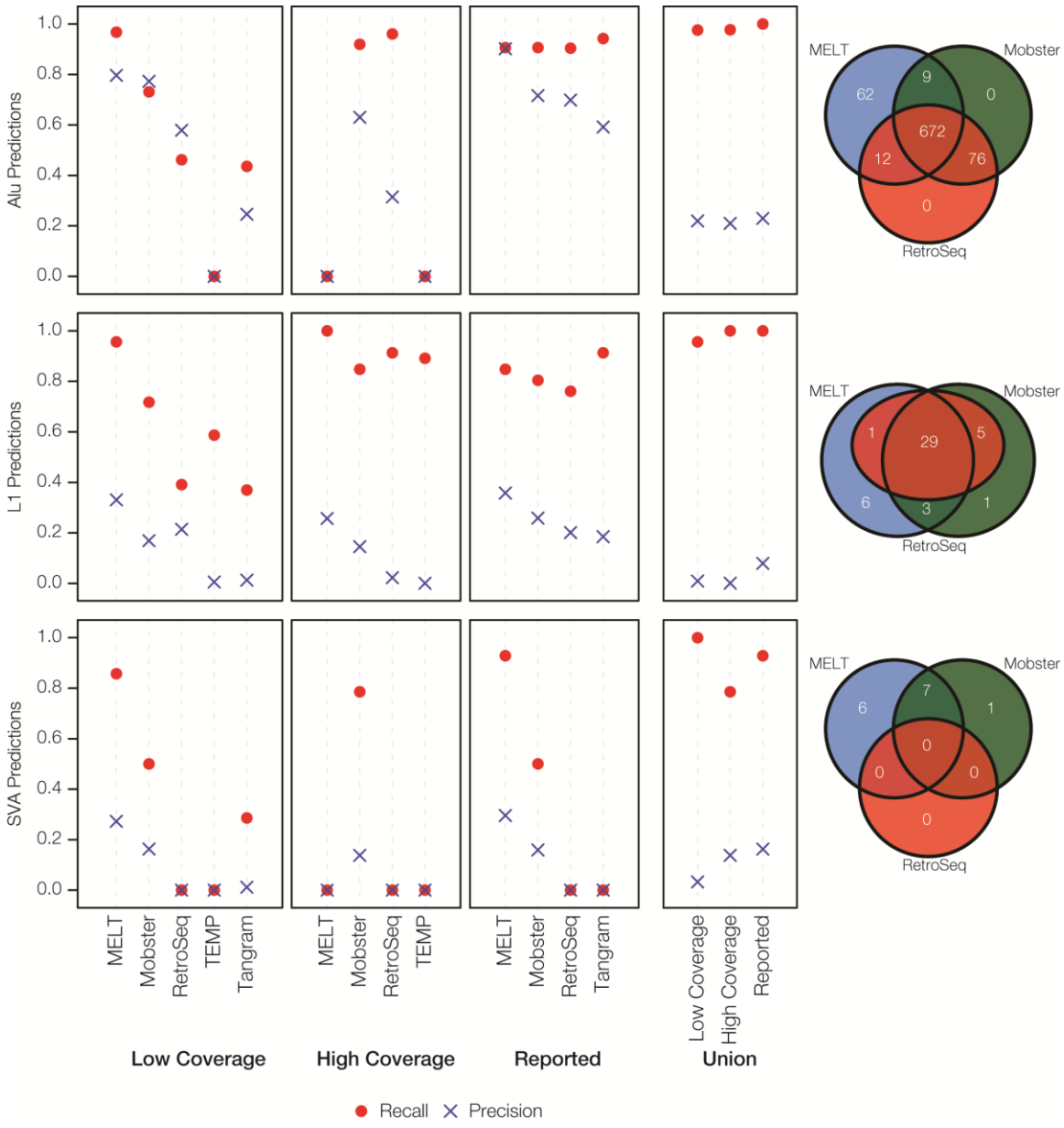

Figure 8 Family-specific polyTE detection tool performance.

TE family-specific Precision and Recall values are shown for the different polyTE detection tools evaluated here for the low (5.7x) and high (95.6x) coverage human genome sequence NA12878. The same parameter values are shown based on previous reports on these tools. The union of predictions made by all tools under consideration is shown for each category. Venn diagrams compare the numbers of unique and shared polyTE insertions reported to have been detected by the three most reliable methods: MELT, Mobster and RetroSeq.

2.4.2 Sequence coverage and tool performance

The low (5.7x) and high (95.6x) coverage data sets described in the previous section represent extreme differences in sequencing depth. We used simulated polyTE insertion data sets across a range of coverages (5x, 10x, 15x, 30x & 50x) in order to more systematically evaluate the effect of sequence depth on the polyTE detection methods evaluated here (Table 4 and Figure 9). The overall performance of the polyTE tools for the simulated data sets is lower than seen for the actual data, indicating that the tools evaluated here work on empirically observed characteristics of polyTE insertions, which cannot be replicated in their entirety via the simulation of *in silico* polyTE data sets. Nevertheless, the relative performance of the tools is very similar to what is seen for the actual data, and it remains stable across the different coverage levels. MELT shows the best overall performance followed by Mobster and then RetroSeq. Recall increases consistently across coverage levels for these three tools, whereas Precision peaks and then flattens out or declines owing to an increase in false positives at higher coverage levels. The overall trend suggests that performance is flattening out or diminishing at ~30x-50x, suggesting a possible coverage limit for these kinds of tools. ITIS gave consistently poor results for these simulated data, whereas TEMP and Tangram failed to make predictions or gave errors.

It should be noted that we also generated a number of additional *in silico* data sets using different simulation parameters than those described for the results reported here. The goal of these additional simulations was to evaluate the effect of different fragment lengths on polyTE detection tools. We evaluated fragment (insert) lengths of 1Kb, 3Kb, 5Kb and 8Kb, which are more typical of mate-pair sequencing technology as opposed to the paired-

end technology used to generate the empirical and simulated data evaluated here. The data sets simulated with longer fragments failed to generated reliable results using any of the tools we evaluated. These results (or lack thereof) underscore the extent to which polyTE detection tools are designed for widely used Illumina paired-end sequencing technology; investigators who wish to use whole genome sequence data for polyTE discovery should be aware of this limitation.

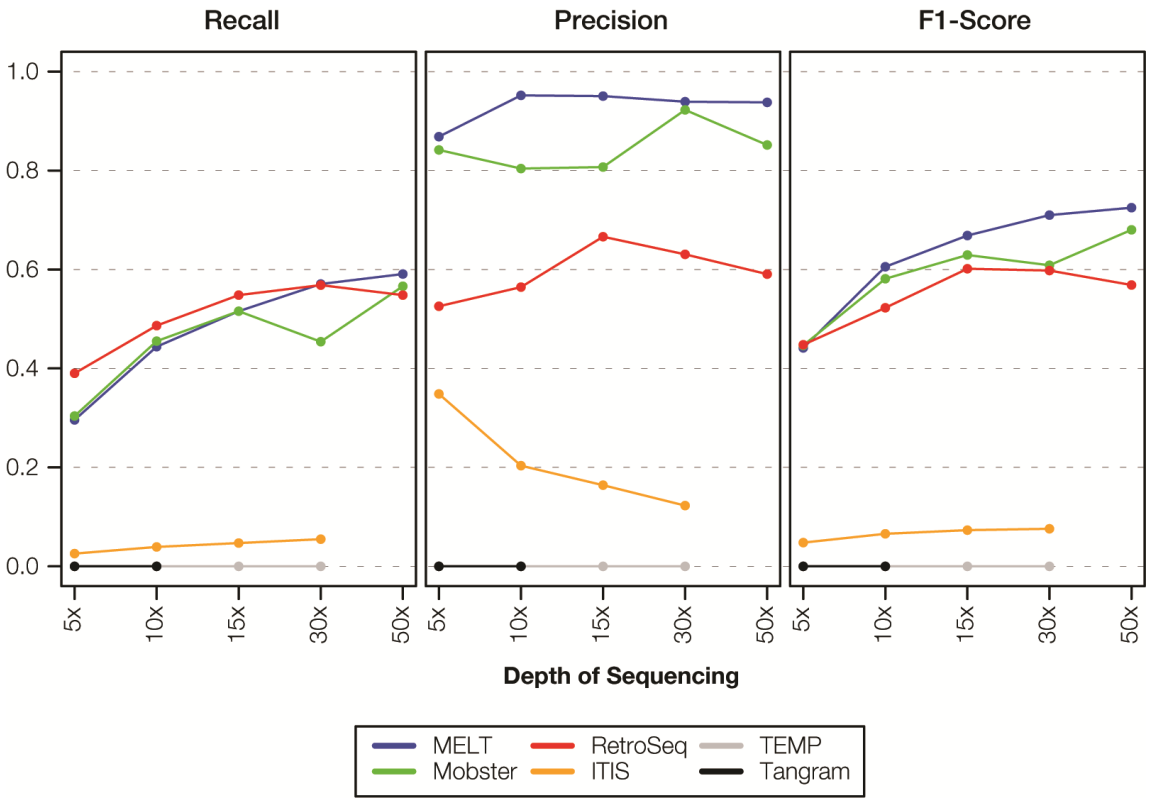

Figure 9 Effect of sequence coverage on polyTE detection tool performance.

Precision, Recall and F1-Scores are shown for the different polyTE detection tools evaluated here across a range of sequence coverages (5x, 10x, 15x, 30x & 50x) from the simulated TE insertion data set.

2.4.3 Runtime parameters

A variety of runtime parameters were measured for the tools on both actual and simulated data as previously described (Table 4). The overall trends are very similar for the actual and simulated data (Figure 10). The programs' runtimes vary over several orders of magnitude and increase in a nearly linear fashion with increasing sequence coverage. The only exception to this trend is seen for ITIS, which has by far the longest runtime and increases much more precipitously with increasing coverage. The CPU time and wall time are closely coupled for most of the tools analyzed here, indicating that the processes executed by the tools are CPU-bound and do not take advantage of parallel execution on multiple cores. ITIS was again the exception to this pattern showing much higher CPU than wall time, consistent with parallel processing on multiple cores. However, this potential advantage is mitigated by its overall long runtime (and poor performance). In addition to its superior performance, MELT is also distinguished by a relatively fast runtime.

Peak memory usage is fairly similar for most of the tools analyzed here and falls well within the range of RAM available for most servers. RetroSeq has an extremely light memory footprint (<1GB RAM) indicating that it can be run on virtually any computer. Results from the percent CPU utilization indicate that most of the tools evaluated here only used one core for most of their runtime, with the exception of ITIS whose percent CPU utilization scales with sequence coverage. In theory, this should yield superior performance, but that was not observed in this case.

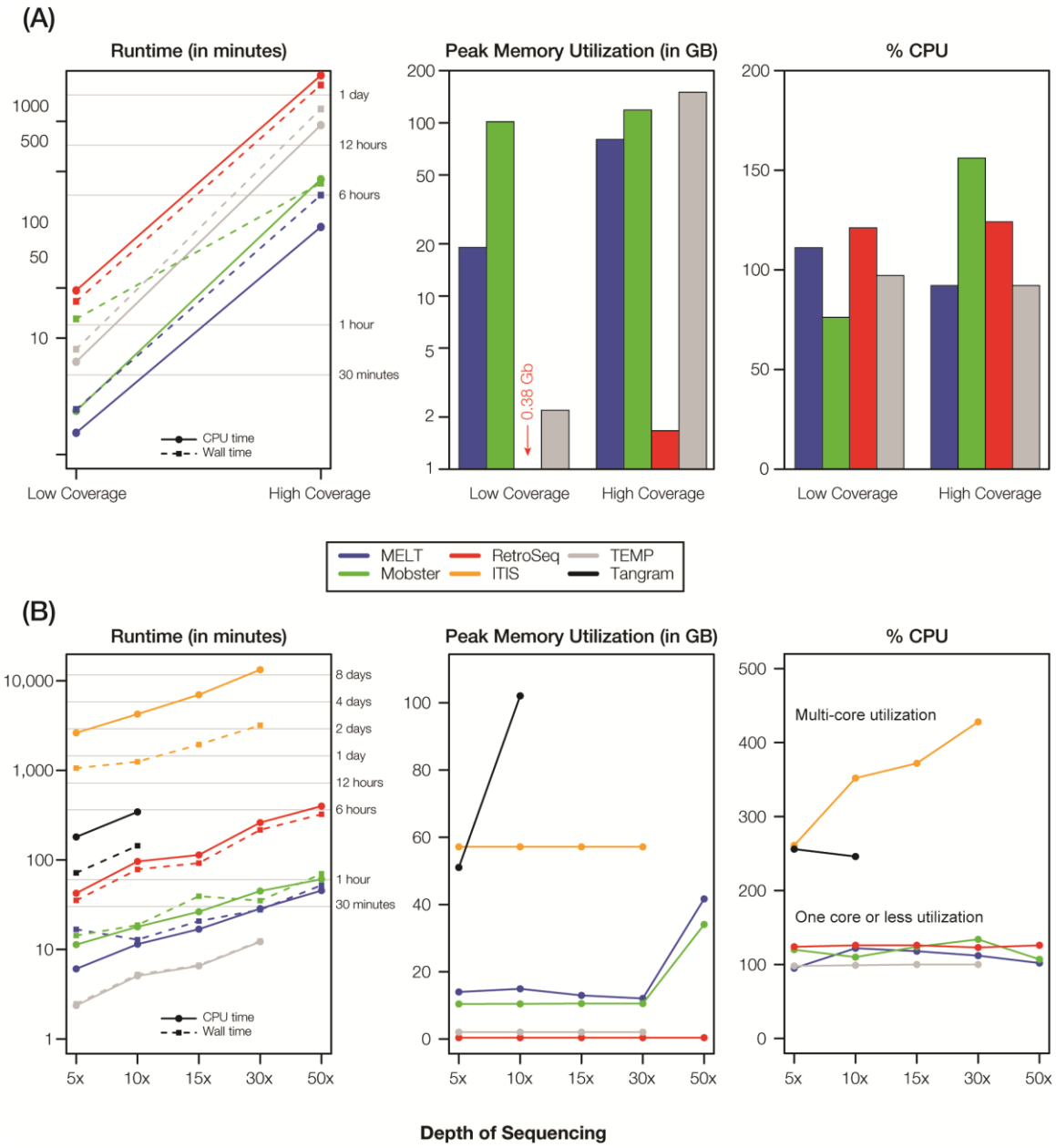

Figure 10 PolyTE detection program runtime parameters.

Runtime, memory and CPU usage are shown for polyTE detection programs run on (A) actual and (B) simulated data sets.

Interestingly, the runtime parameters do not seem to be affected by the choice of programming language used by developers of the different tools, as is commonly believed by programmers. For example, Tangram is written in C++ and thus should in principle be much faster and more efficient than the other programs written in Perl; this did not prove to be the case. On the other hand, RetroSeq is written in Perl but has the lowest memory footprint, contrary to what may be expected. Clearly, the programming language of choice is less relevant than the algorithm design principles employed by these programs. This may be a truism, but it also may point to the opportunity for substantial future improvement in the design of these tools.

2.5 Additional notes for users and developers

We provide detailed notes on the installation and use of the benchmarked polyTE detection programs in the Appendix A. Here, we provide some more general notes on practical issues that users of these programs should be aware of, along with possible suggestions for developers related to these same issues.

1. ***Installation of dependencies:*** Some of the tools require that users install dependencies from third party developers that are not bundled with the tool. This seemingly trivial requirement can be prove to be quite challenging for both relatively naïve users and in the case where the dependency version changes affect program output. We recommend that developers bundle all dependencies with their polyTE detection software.
1. ***Parameter choice:*** Some of the tools require that users provide a number of parameters, many of which could be easily calculated from the input data sets.

We recommend that developers consider automatic parameter calculation from input data, where possible, to allow for ease of use and improved performance.

2. ***Input reference databases:*** Most of the tools have specific formatting requirements for the databases of TE consensus sequences and/or coordinates that users are required to provide. Generation of such tool-specific reference databases is time consuming and potentially error prone. We recommend that developers provide pre-formatted reference databases for human and model organisms to facilitate accurate and ready use of their tools.
3. ***Filtering TE predictions:*** Some of the tools have criteria by which users should filter the automatically generated output of TE predictions (*e.g.*, number of reads that support predictions). Lack of guidance as to specific filtering criteria leads to numerous false positives. We recommend that developers provide the option for filtering based on parameters derived from the input data set (see point #2 above).
4. ***VCF output:*** Output formats vary among the tools evaluated here. Variant call format (VCF) is a generally accepted and widely used format for variant representation. Availability of VCF output would allow for ease of interpretation and better integration with downstream analysis tools.

2.6 Conclusions and future prospects

The polyTE detection tool MELT shows consistently superior performance on the human genome sequence data (actual and simulated) analyzed here. The only exception to this trend was seen for the very high coverage data set, where MELT failed to predict Alu and SVA insertions. The superior performance of MELT may be related to the fact that it was the program used by the 1KG Project Structural Variation Group to make predictions on

the same sample (NA12878) that was used for validation purposes here. In addition, MELT takes advantage of prior information on the known locations of human polyTE insertions. Despite these caveats, or perhaps owing in part to the additional information gained during the development process of the tool, MELT is currently the best choice for the detection of human polyTE insertions.

In our hands, Mobster and Retroseq were slightly less reliable options for human polyTE detection. These tools showed consistent performance across the data sets analyzed here, and they were both relatively easy to install and run. RetroSeq is further distinguished by a particularly light computational footprint that makes it useable on virtually any computer. None of the other tools benchmarked here are currently recommended for the detection of human polyTE insertions. It is formally possible that some of the more poorly performing tools may in fact work well in the hands of their developers, and that the performance metrics reported here reflect the fact that we were unable to get them to work correctly. However, whenever we had problems with tool use, we made efforts to thoroughly review the documentation, verify the input and reference files, vary usage parameters and change the dependency versions. When none of this worked, we contacted the developers directly for their feedback. Thus, we made extensive efforts to get the tools to work, and our ability (or lack thereof) to do so can be considered as an important source of information for potential tool users and developers. It may also be the case that some of the tools evaluated here, such as Tangram, are no longer actively supported and represent a stage in the ongoing development of polyTE detection algorithms.

Another caveat is that some of these tools were developed for other model organisms. For example, ITIS was developed for the plant *Medicago trunculata*, and T-lex2 was developed

for the analysis of Drosophila sequence data. It is possible that their relatively poor performance on human data sets reflects the fact that they are better tuned to the TEs and genomic sequence context of their respective organisms. In addition

It is also worth noting that the union of predictions made by all the methods under consideration always yields higher Recall than any single method (Figure 7 and Figure 8). Thus, the polyTE detection tools evaluated here may be considered to be complementary. Of course, combining results of all methods for any given set of predictions yields numerous false positives. Nevertheless, a careful combined analysis - using MELT, Mobster and RetroSeq for example - with some kind of majority rule criterion and/or careful manual (visual) inspection of read mapping results may provide for the optimal polyTE detection.

Despite the fact that we ran all of these tools on a high performance server with substantial memory and processing power, several of the tools ran for an extremely long time and/or failed to produce output. In some cases, higher coverage, which should in principle allow for improved performance, severely impeded the programs execution. A number of these tools have been developed by genome analysis consortia and/or as part of large-scale sequencing efforts, which are likely to have substantial computational resources at their disposal. But in order for these tools to be widely adopted by the research community, a concerted effort will have to be made to ensure that they are both user-friendly and scalable. This suggests an excellent opportunity for developers to create algorithms that are more computationally efficient and thereby more widely accessible to the research community. In short, there is still a lot room for development in the area of polyTE detection.

Finally, it is important to note that many of the large-scale human genome projects underway will continue to use short read sequencing technology, Illumina in particular, which is by far the current industry leader for re-sequencing. Accordingly, the use of the kinds of polyTE detection tools evaluated here will remain critical for the characterization of TE-generated genetic variation. However, the era of single molecule sequencing is very much underway, and the long sequence reads generated by technologies such as PacBio and Oxford Nanopore would render these short read computational techniques irrelevant. But it is currently unclear whether, and the extent to which, such long sequence read technologies may eventually supplant Illumina for human genome re-sequencing.

CHAPTER 3. TRANSPOSABLE ELEMENT POLYMORPHISMS RECAPITULATE HUMAN EVOLUTION

3.1 Abstract

The human genome contains several active families of transposable elements (TE): Alu, L1 and SVA. Germline transposition of these elements can lead to polymorphic TE (polyTE) loci that differ between individuals with respect to the presence/absence of TE insertions. Limited sets of such polyTE loci have proven to be useful as markers of ancestry in human population genetic studies, but until this time it has not been possible to analyze the full genomic complement of TE polymorphisms in this way. For the first time here, we have performed a human population genetic analysis based on a genome-wide polyTE data set consisting of 16,192 loci genotyped in 2,504 individuals across 26 human populations. PolyTEs are found at very low frequencies, >93% of loci show <5% allele frequency, consistent with the deleteriousness of TE insertions. Nevertheless, polyTEs do show substantial geographic differentiation, with numerous group-specific polymorphic insertions. African populations have the highest numbers of polyTEs and show the highest levels of polyTE genetic diversity; Alu is the most numerous and the most diverse polyTE family. PolyTE genotypes were used to compute allele sharing distances between individuals and to relate them within and between human populations. Populations and continental groups show high coherence based on individuals' polyTE genotypes, and human evolutionary relationships revealed by these genotypes are consistent with those seen for SNP-based genetic distances. The patterns of genetic diversity encoded by TE polymorphisms recapitulate broad patterns of human evolution and migration over the last

60-100,000 years. The utility of polyTEs as ancestry informative markers is further underscored by their ability to accurately predict both ancestry and admixture at the continental level. A genome-wide list of polyTE loci, along with their population group-specific allele frequencies and F_{ST} values, is provided as a resource for investigators who wish to develop panels of TE-based ancestry markers. The genetic diversity represented by TE polymorphisms reflects known patterns of human evolution, and ensembles of polyTE loci are suitable for both ancestry and admixture analyses. The patterns of polyTE allelic diversity suggest the possibility that there may be a connection between TE-based genetic divergence and population-specific phenotypic differences.

3.2 Background

Much of the human genome sequence, anywhere from ~50-70% depending on estimates [2, 19], is derived from transposable elements (TE). The vast majority of TE-derived sequences in the genome are remnants of ancient insertion events, which are no longer capable of transposition. Nevertheless, there remain a few families of actively transposing human TEs [3]; the active families of human TEs include Alu [29, 30], L1 [27, 28] and SVA [31, 32] elements. Alu elements are 7SL RNA-derived short interspersed nuclear elements (SINEs) [38, 39], L1s are a family of long interspersed nuclear elements (LINEs) [33, 34], and SVA elements are composite TEs that are made up of human endogenous retrovirus sequence, simple sequence repeats and Alu sequence [40, 41]. All three of these active families of human TEs are retrotransposons that transpose via reverse transcription of an RNA intermediate. L1s are autonomous retrotransposons that encode the enzymatic

machinery necessary to catalyze their own retrotransposition [35], whereas Alu and SVA elements are transposed in *trans* by the L1 machinery [36, 37].

If members of these active TE families transpose in the germline, they can create novel insertions that are capable of being inherited, thereby generating human-specific polymorphisms. Such polymorphic TE (polyTE) insertion sites have been shown to be valuable genetic markers for studies of human ancestry and evolution. PolyTEs provide a number of advantages for such population genetic studies [3, 17]. First, the presence of a polyTE insertion site shared by two or more individuals nearly always represents identity by descent [17, 121]. This is because there are so many possible insertion sites genome-wide, and transposition rates are so low, that the probability of independent insertion at the same site in two individuals is negligible. Second, since newly inserted TEs rarely undergo deletion they are highly stable polymorphisms. These two characteristics underscore the fact that polyTE markers are completely free of homoplasies, *i.e.* identical states that do not represent shared ancestry, which are far more common for single nucleotide polymorphisms (SNPs). Another useful feature of polyTEs for population genetic studies is the fact that the ancestral state of polyTE loci is known to be absence of the insertion [46, 47]. Finally, polyTEs are practically useful markers since they can be rapidly and accurately typed via PCR-based assays.

A number of previous studies have leveraged TE polymorphisms for the analysis of human ancestry and evolution [3, 17, 37, 46-52]. Most of these studies have focused on Alu elements; there have been far fewer human population genetic studies using L1 markers and to our knowledge no such studies using polymorphic SVA elements. Alus are particularly advantageous for these types of studies because their small size allows them to

be readily PCR amplified; furthermore, both the presence and absence of Alu insertions can yield amplification products from a single PCR. Ancestry studies that use TE polymorphisms have relied on a number of selection criteria in order to try and define the most useful polyTE loci for human population differentiation. For instance, polyTE loci have often been identified via literature surveys of specific gene mutations caused by TE insertions. Analysis of the human genome sequence has also been used to identify intact members of the youngest (*i.e.* recently active) subfamilies of Alus and L1s in order to try and predict potentially mobile sequences. Once potential polyTE marker loci are chosen using these methods, they need to be empirically evaluated with respect to their levels of polymorphism within and between populations. These approaches, while somewhat *ad hoc* and laborious, have in fact proven to be useful for the identification of polyTE loci that serve as ancestry informative markers (AIMs).

The most recent data release from the 1000 Genome Project (Phase3 November 2014) includes, for the first time, a comprehensive genome-wide data set of polyTE sites. There are a total of 16,192 such polyTE loci reported for 2,504 individuals across 26 human populations. These newly available data provide an unprecedented level of depth and resolution for polyTE-based studies of human ancestry and evolution. With these data, it is now possible to evaluate the relationship between TE polymorphism and human evolution in a systematic and unbiased way. In addition, individual polyTE loci genome-wide can be evaluated with respect to their utility as AIMs as well as their applicability to ancestry studies for specific population groups. Such an analysis could provide a useful resource for investigators interested in conducting their own targeted studies on specific populations. With such a comprehensive, genome-wide polyTE data set, it is also possible

to evaluate the marker utility of previously under-utilized L1 and SVA sequences. For this study, we have conducted a genome-wide population genetic analysis of human TE polymorphisms in order to address precisely these kinds of issues. This work represents the most comprehensive study of human polyTEs to date.

3.3 Results

3.3.1 Human population genomics of polyTEs

There are three families of polymorphic transposable elements (polyTEs) that show variation in presence/absence patterns at individual insertion sites across human genome sequences; these are Alu (SINE), L1 (LINE) and chimeric SVA elements. The Phase3 data release (November 2014) of the 1000 Genomes Project provides the most complete catalog of human transposable element insertion site polymorphisms available to date. Presence/absence genotypes for these human polyTEs are available for 2,504 individuals from 26 human populations across 16,192 genomic sites.

Table 5 Human populations analyzed in this study.

Populations are organized into five continental groups, and the number of individuals in each population is shown. The same population-specific color codes are used throughout the manuscript.

	Color	Short	Full Description	*n*
African (*n*=504)		ESN	Esan in Nigeria	99
		GWD	Gambian in Western Division, The Gambia	113
		LWK	Luhya in Webuye, Kenya	99
		MSL	Mende in Sierra Leone	85
		YRI	Yoruba in Ibadan, Nigeria	108
Asian (*n*=504)		CDX	Chinese Dai in Xishuangbanna, China	93
		CHB	Han Chinese in Bejing, China	103
		CHS	Southern Han Chinese, China	105
		JPT	Japanese in Tokyo, Japan	104
		KHV	Kinh in Ho Chi Minh City, Vietnam	99
European (*n*=503)		CEU	Utah residents with NW European ancestry	99
		FIN	Finnish in Finland	99
		GBR	British in England and Scotland	91
		IBS	Iberian populations in Spain	107
		TSI	Toscani in Italy	107
Indian (*n*=489)		BEB	Bengali in Bangladesh	86
		GIH	Gujarati Indian in Houston,TX	103
		ITU	Indian Telugu in the UK	102
		PJL	Punjabi in Lahore,Pakistan	96
		STU	Sri Lankan Tamil in the UK	102
American (*n*=504)		ACB	African Caribbean in Barbados	96
		ASW	African Ancestry in Southwest US	61
		CLM	Colombian in Medellin, Colombia	94
		MXL	Mexican Ancestry in Los Angeles, California	64
		PEL	Peruvian in Lima, Peru	85
		PUR	Puerto Rican in Puerto Rico	104

We characterized the frequencies and distributions of human polyTEs for the 26 populations organized into 5 continental groups: African, Asian, European, Indian and American (Table 5). The vast majority of human polyTEs are found at low frequencies

within and between human populations; 15,141 (93.5%) of polyTE loci show <5% overall allele frequencies (Figure 11A). Nevertheless, there is substantial variability of individual polyTE allele frequencies among populations from different continental groups (Figure 11B). Accordingly, there are higher numbers of polyTEs with continental group-specific allele frequencies >5% (Figure 11C), and numerous individual polyTE loci are exclusively present within a single continental group (Figure 11D). On average, ~25% of individual polyTE loci are exclusive to a specific continental group. These results are consistent with the possibility that polyTE genotypes may serve as useful markers of genomic ancestry. Results of the same analyses are shown for individual polyTEs families in Figure 22. Alu is by far the most abundant family of polyTEs followed by L1 and SVA. All three polyTE families show similar levels of continental group-specific insertions.

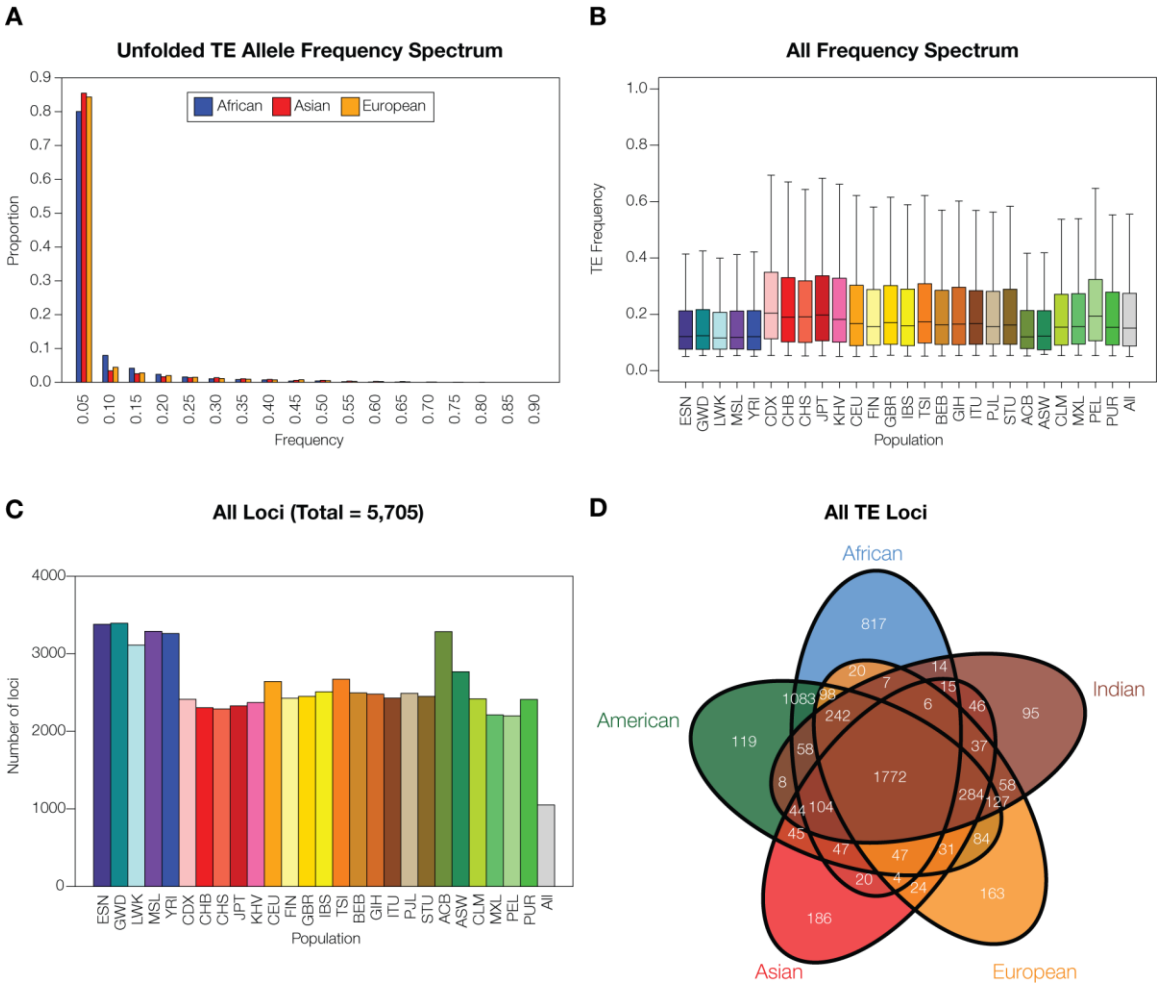

Figure 11 Distribution of polymorphic transposable element (polyTE) loci among human populations.

Populations are organized into five continental groups (see Table 5): African (blue), Asian (red), European (gold), Indian (brown) and American (green). (A) Unfolded polyTE allele frequency spectrum for the three ancestral (non-admixed) continental groups: African, Asian and European. (B) Boxplot polyTE allele frequency distributions for TE insertions present at >5% frequency within individual populations. (C) Numbers of polyTE loci at >5% frequency that are shared or exclusive among continental groups. (D) Numbers of polyTE loci at >5% frequency among the different populations.

PolyTE genotypes were analyzed in order to evaluate the polyTE genetic diversity levels for different continental groups and for different TE families. To do this, presence/absence

patterns at all polyTE loci were used to genotype individual human genomes and pairwise allele sharing distances between individuals were computed based on these polyTE genotypes (see Materials and Methods). African populations have the highest levels of polyTE genetic diversity and Asian populations show the lowest diversity (Figure 12A). These data are similar to what has been in previous studies of polyTEs [52] and for SNP-based genetic diversity [103]. All of the differences in median genetic diversity levels between pairs of population groups are highly statistically significant ($0 \leq P \leq 8.5 \times 10^{-56}$ Wilcoxon ranked sum test). African populations also have the highest levels of variation in polyTE genetic diversity for any of the non-admixed groups, consistent with human origins in Africa and the bottleneck experienced by other population groups during their migrations out of Africa [122, 123]. The overall effect of recent admixture in the Americas is revealed by the broad distribution of polyTE genetic diversity among the American populations, and African admixture among these same populations probably accounts for the fact that this group has the second highest level of median diversity seen for all continental groups (Figure 12A). For polyTE families, Alu has the highest diversity followed by SVA and L1 (Figure 12B). The relative levels of continental group polyTE genetic diversity are the same for all three families of polyTEs (Figure 12C-D).

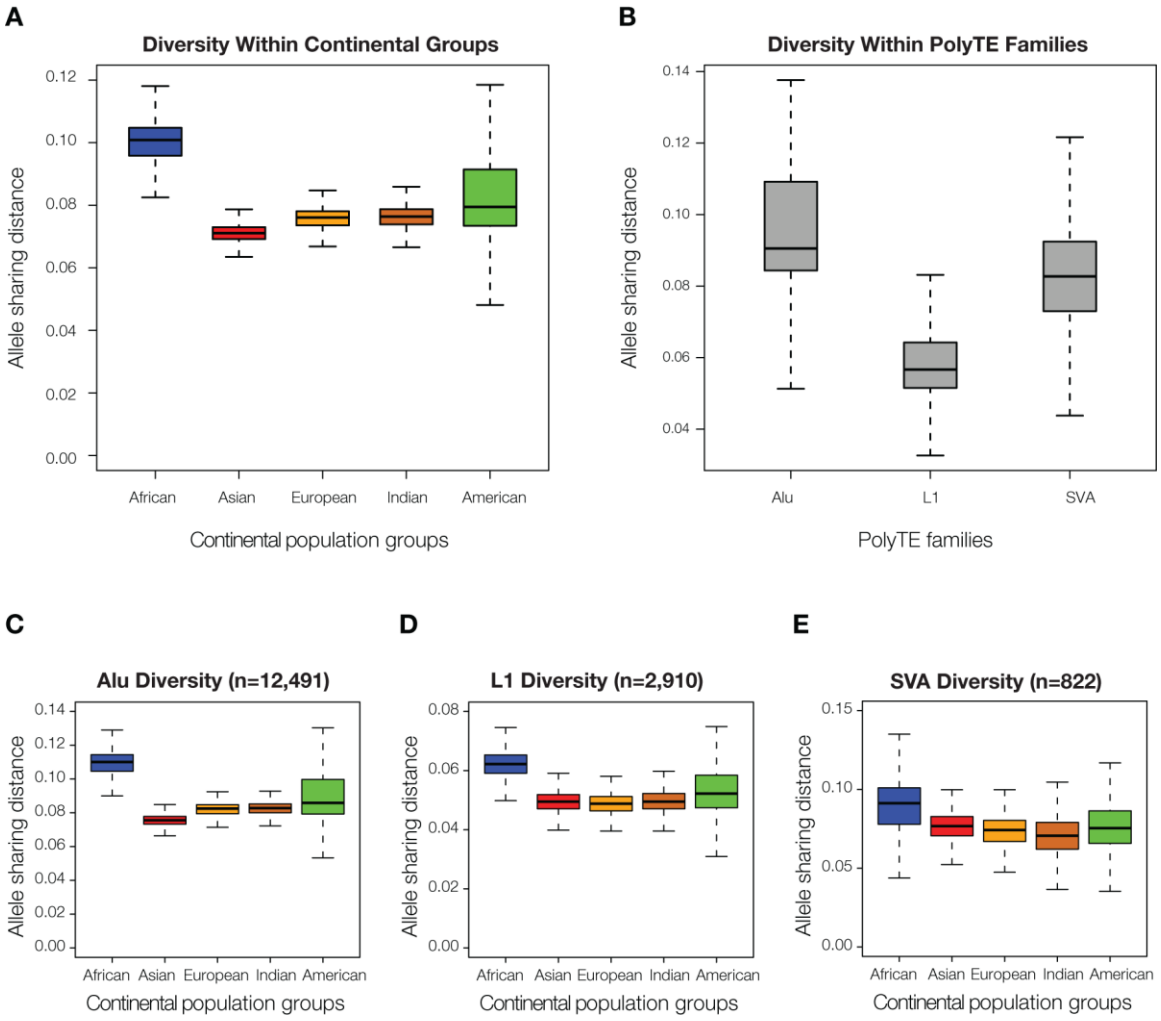

Figure 12 PolyTE genetic diversity levels.

(A) Distributions of overall polyTE genotype-based allele sharing distances are shown for the five continental groups (see Table 5): African (blue), Asian (red), European (gold), Indian (brown) and American (green). (B) Distributions of polyTE genotype-based allele sharing distances are shown for separately Alu, L1 and SVA. (C-E) TE family-specific distributions of polyTE genotype-based allele sharing distances are shown for separately Alu, L1 and SVA.

3.3.2 Human evolutionary relationships based on polyTEs

The distributions of polyTE genotypes among individuals were analyzed in an effort to reconstruct the evolutionary relationships among human individuals and populations. To do this, PolyTE genotype allele sharing distances were used to generate multi-dimensional scaling (MDS) plots showing the genetic relationships among all individuals (Figure 13A) and the average genetic relationships between individual populations (Figure 13B). Phylogenetic reconstruction was also used to show the average polyTE genotype-based relationships between populations (Figure 13C). The evolutionary relationships revealed by this analysis are entirely consistent with previous analyses based on individual nucleotide level variation assessed via SNP-based genotypes [124], and very similar to what has previously been seen based on Alu polymorphisms [48]. African, Asian and European continental groups represent the three poles of human genomic variation with the more ancient admixed Indian group and more recent admixed American group in between. In the phylogenetic analysis, the African populations are the most basal with the European and Asian populations being derived.

One of the advantages of using TE polymorphisms for ancestry inference is that the ancestral state for any polyTE loci can be confidently taken to be the absence of an insertion [46, 47]. This property allows for the creation of a hypothetical ancestral genome characterized by the absence of insertions across all polyTE loci. When such a hypothetical ancestor is included in the polyTE-based reconstruction of human evolutionary relationships, it maps near the center of the MDS plots closer to the African populations (Figure 13A and 13B), and it maps closest to the root of the phylogeny between the African

and non-African lineages (Figure 13C). These results confirm that polyTE insertions are derived allelic states.

For the most part, there is high coherence of polyTE genotypes within both individual populations and for continental groups. The only exception seen is for the admixed American continental group, which has two distinct subgroups, a Latino subgroup (PEL, MXL, CLM and PUR) with primarily European and Asian admixture and an African-American subgroup (ACB and ASW) with primarily African and European admixture (Figure 13D). The relative admixture levels seen for these populations are consistent with previous nucleotide level SNP-based analysis [125, 126]. The apparent Asian admixture of the Latino subgroup reflects Native American ancestry owing to the fact that Native Americans are relatively recently derived from East Asian populations [127]. As there are no Native American samples in the 1000 Genomes Project Data [103, 128], the East Asian genome sequences appear as most closely related to the Latino subgroup. CLM and PUR show relatively higher levels of European, and to a lesser extent African, admixture than seen for PEL and MXL (Figure 13D). We also attempted to infer Native American ancestry in admixed American populations by imputing polyTE genotypes for Native American populations from the Human Genome Diversity Project based on the 1000 Genome Project imputation panels. The ancestry contribution fractions for admixed American individuals are highly correlated between the observed Asian polyTE genotypes and the imputed Native American polyTE genotypes (Figure 23).

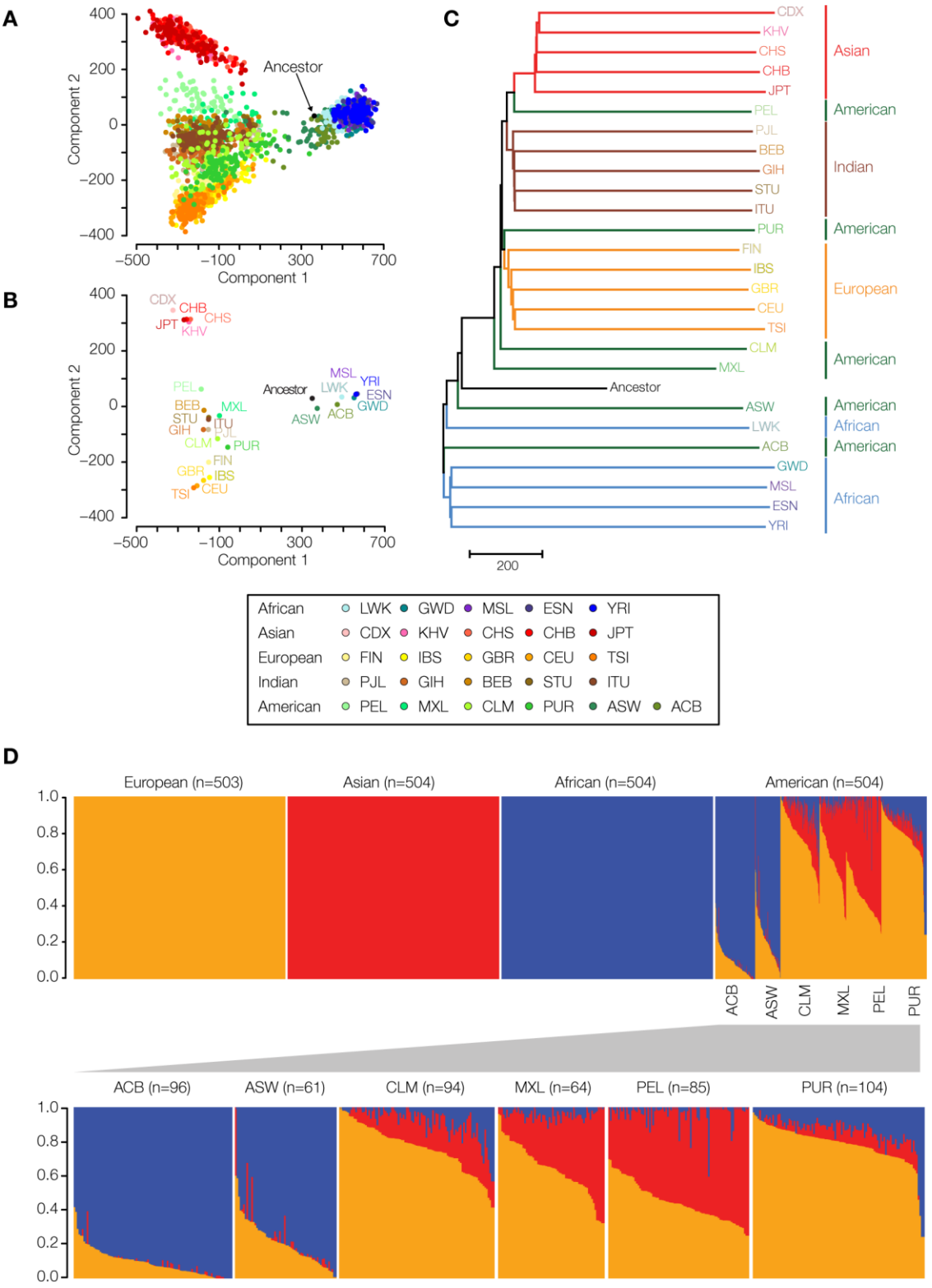

Figure 13 Evolutionary relationships among human populations based on polyTE genotypes.

Populations are color coded as shown in the figure legend. (A) Multi-dimensional scaling (MDS) plot showing polyTE genotype-based distances among 2,504 individuals from 26 human populations. (B) The same polyTE genotype MDS plot showing population average distances. (C) Phylogenetic tree based on average polyTE allele sharing distances between human populations. (D) polyTE genotype-based continental ancestry contribution fractions for individuals from non-admixed ancestral (European, Asian and African) and admixed (American) human populations. An expanded view of the ancestry fractions is shown for the admixed American populations.

Results of the same analyses are shown for individual polyTEs families in Figures 24-26. While the results are highly concordant for all three polyTE families, Alu ployTEs show the highest levels of resolution for human evolutionary relationships owing to the far higher number of polymorphic Alu insertions available for analysis. Nevertheless, L1 and SVA elements also show the ability to differentiate human populations and continental groups suggesting that these previously under-utilized polyTEs may also serve as useful ancestry markers.

3.3.3 Ancestry prediction with polyTEs

Having established the overall ability of polyTE-based genotype analysis to capture known evolutionary relationships among human populations, we evaluated the ability of individual of polyTE loci to serve as useful markers for ancestry inference. To do this, levels of population differentiation for individual polyTE loci were assessed using the fixation index F_{ST} and the absolute allele frequency differences δ (see Materials and Methods). PolyTE loci-based F_{ST} and δ distributions were computed for three-way comparisons between non-admixed continental groups (African, Asian and European) and

for five-way comparisons between individual populations within the same non-admixed continental group (Figures 27 and 28). As can be expected, individual polyTE loci show substantially higher levels of population differentiation (*i.e.* higher F_{ST} and δ values) for the between compared to the within continental group comparisons. This is consistent with the overall ability of polyTE genotypes to better distinguish between continental groups (Figure 13) than within continental groups (Figure 29). The same pattern has been observed for SNP-based AIMs [129]. Nevertheless, polyTE loci are able to provide some level of resolution for even closely related populations within continental groups. A comprehensive list of human polyTE loci along with their allele frequencies and F_{ST} and δ values, within and between populations, are provided in Table 11 so that investigators can choose loci of interest as potential ancestry markers.

Interestingly, the overall levels of polyTE-based F_{ST} are fairly low even for the between continental group comparison (Figure 27). F_{ST} levels $\geq$0.4 have previously been taken to indicate that a nucleotide SNP can serve as a useful ancestry informative marker (AIM) [129, 130]. There are no individual polyTE loci that conform to this AIM criteria; 0.39 is the highest polyTE F_{ST} value. This can be attributed to the overall low frequency of polymorphic TE insertions seen here (Figure 11A) since low levels of within-group polyTE allele frequency will depress F_{ST} levels owing to high levels of within group heterozygosity. The values of δ appear to be somewhat more sensitive for the characterization of individual polyTE AIMs. Several different δ value thresholds have been proposed for AIM characterization over the years [129]: 0.3, 0.4 and 0.5. There are 371 (0.3), 79 (0.4) and 9 (0.5) polyTE loci with continental δ values that exceed these thresholds. Thus, individual polyTE loci appear to have moderate ability to differentiate

human populations, whereas ensembles of polyTE loci can be used effectively to distinguish more closely and distantly related populations.

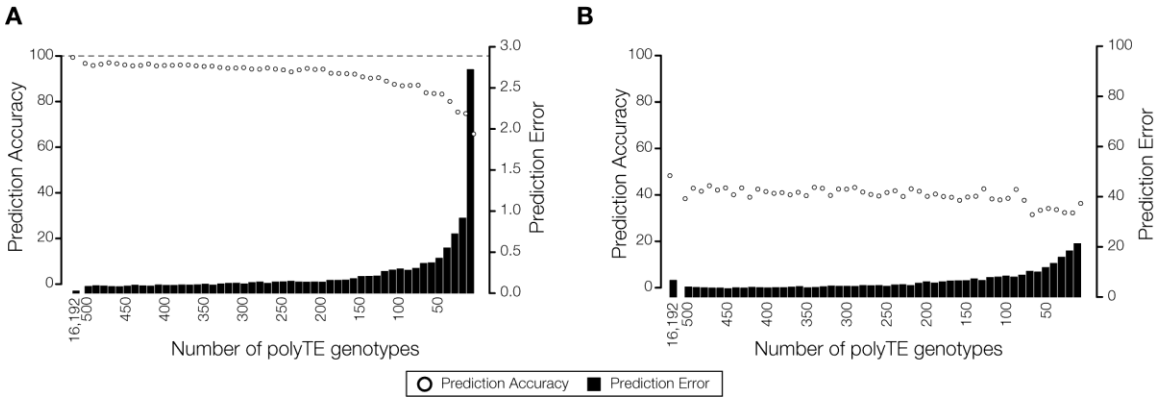

Figure 14 Ancestry predictions using polyTE genotypes.

Relationship between the number of polyTE loci used to genotype individuals and the ancestry prediction accuracy for (A) continental population group comparisons (between African, Aisan and European) and (B) sub-continental population comparisons (European).

In light of the ability of individual polyTEs genotypes and overall polyTE genotype patterns to differentiate human populations, we attempted to identify the smallest set of polyTE loci needed to accurately predict human ancestry. The accuracy of ancestry prediction was assessed for both non-admixed continental groups (African, Asian and European) and for individual populations within the African continental group. To do this for each comparison, the top 500 ancestry informative polyTE loci were ranked according to their F_{ST} levels and prediction accuracy was computed for sets of polyTE loci of sequentially decreasing size, going from 500 to 10 in steps of 10 (Figure 14). Two measures of ancestry prediction, accuracy and error, were measured for each set of polyTE loci using the approach described in the Materials and Methods. When all polyTE loci are

used, continental group ancestry prediction approaches 100% accuracy with <1% error. As the number of polyTE loci used for ancestry prediction is steadily decreased from 500, the accuracy declines and the error increases. However, the changes in accuracy and error are relatively slight. For the top 100 polyTE loci, ancestry prediction is 86.9% accurate with 0.3% error. The smallest set of 10 polyTE loci yields 65.8% accuracy and 2.7% error. These results are similar to previous report [52] that evaluated the minimum number of polymorphic Alu loci (~50) that would yield accurate genetic distances between human populations.

A similar approach was taken to evaluate the utility of polyTE genotypes for ancestry prediction within continental groups. Consistent with what is observed for the within continental group F_{ST} values (Figure 27), polyTE genotypes have less power to discriminate ancestry for closely related populations from the same continental group (Figure 14B). For the African populations, individual genotypes based on the entire set of polyTE loci yield an ancestry prediction accuracy of 48.3% and an error of 6.7%. Since there are five African populations, a random predictor would yield 20% accuracy. Thus, the accuracy achieved by polyTE loci, while relatively low, is 2.4x greater than expected by chance alone. Accuracy does not change greatly with decreasing numbers of polyTE loci. 100 polyTE loci yields accuracy of 38.5%, and the accuracy for 10 polyTE loci is 36.3%. The error rate of prediction does steadily increase to 8.4% for 100 polyTE loci and 21.3% for 10 polyTE loci.

3.3.4 Admixture prediction with polyTEs

Having established the utility of small sets of polyTE loci to make ancestry inferences for non-admixed groups, we wished to similarly evaluate the ability of polyTE loci sets to allow for inferences about continental ancestry contributions to admixed populations. To do this, ancestral contributions from African and European populations to the admixed ASW American population were evaluated using sets of polyTE loci of decreasing size in a similar way as was done for ancestry prediction in non-admixed populations. In the case of admixture, prediction error levels were measured by comparing the ancestral admixture components computed from the entire set of 16,192 polyTE loci to those computed from the smaller polyTE loci sets (see Materials and Methods). As with ancestry prediction, error levels steadily increase with the use of decreasing numbers of polyTE loci (Figure 15A). However, slightly larger numbers of polyTE loci are required to keep admixture inference error levels low; the use of 10 polyTE loci yields 3.4% error, whereas a set of 50 polyTE loci reduces the error to 2.2%. There is strong agreement in the results of continental ancestry contributions for this admixed population between analyses conducted with all polyTEs versus the top 50 polyTEs (r=0.62; Figure 15B).

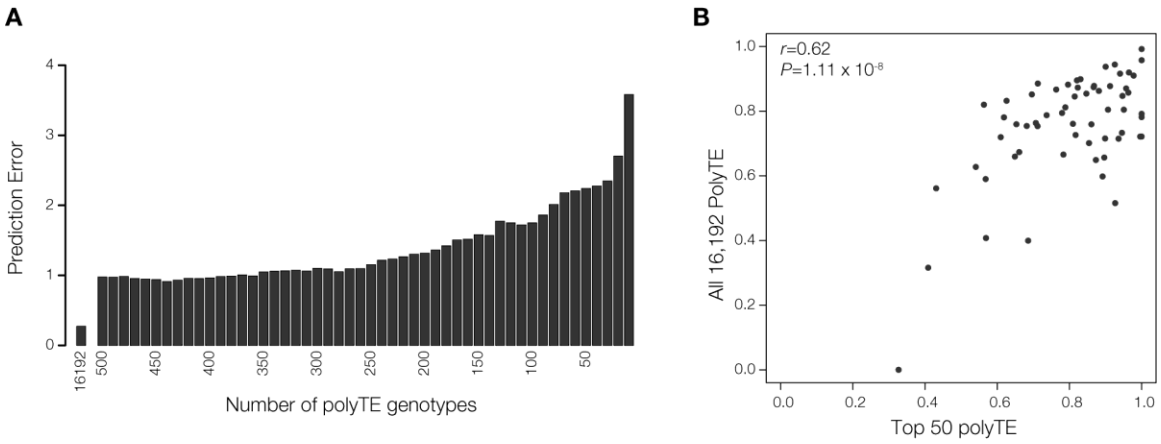

Figure 15 Admixture predictions using polyTE genotypes.

(A) Relationship between the number of polyTE loci used to genotype individuals and admixture prediction accuracy for the ASW population. (B) Comparison of individual admixture proportions calculated using all available polyTE genotypes versus a minimal polyTE genotype set with 50 loci.

3.4 Discussion

3.4.1 Human ancestry and admixture from polyTEs

Our analysis of a genome-wide set of human polyTE genotypes indicates that TE polymorphism patterns recapitulate the pattern of human evolution and migration over the last 60-100,000 years (Figure 13 and Figures 24-26). While polyTEs considered as an ensemble provide substantial resolution for inferring ancestry and human relationships, individual polyTE loci show moderate population differentiation levels (Figure 27 and 28). This can be attributed to the fact that individual polyTE loci tend to be found at low allele frequencies (Figure 11A). However, these same low frequency loci do show high levels of geographic differentiation, *i.e.* many of them are continental group or population specific (Figure 11B). Therefore, when a relatively small set of these low frequency but highly

geographically differentiated polyTE loci are used together, they do in fact provide substantial resolution for evolutionary analysis as well as ancestry and admixture inference (Figures 14 and 15).

These results have important implications for the study of human evolution, ancestry and admixture by smaller labs that may not have access to the same level of resources as larger consortia or genome centers since analysis of a small set of polyTE loci (10-50 depending on the application) can prove to be quite informative. Given the size range of TEs insertions, in particular for Alus which are the most numerous family of polyTEs, element presence/absence patterns can be accurately characterized in a cost-effective way using (multiplex) PCR-based techniques. Protocols for PCR-based analysis of polyTEs are well established in a number of labs. The results of this study can be used to help investigators choose the specific TE loci of interest for their own evolutionary studies (see Table 11 for a list of genomic locations of polyTEs and their allele frequencies and F_{ST} values).

Despite the overall utility of polyTEs as ancestry markers, results from this study suggest that they are not likely to be good markers for mapping by admixture linkage disequilibrium (MALD or admixture mapping) studies [131, 132]. These studies rely on detailed locus-specific assignments of ancestry across the genome in admixed individuals. In order to achieve this level of resolution, thousands of markers are needed and individual markers should have high levels of population differentiation (as measured by F_{ST} or other related metrics) [129] . Thus, SNPs would seem to remain the best choice of AIMs for MALD (admixture mapping) studies.

3.4.2 Deleteriousness and selection on polyTE insertions

Our initial analysis of human polyTEs within and between populations revealed that TE insertion polymorphisms are found at very low frequencies (Figure 11A). This is consistent with the overall deleteriousness of TE insertions and accordingly their removal by purifying selection. The elimination of polyTEs by purifying selection is also underscored by the fact that polyTEs are vastly under-represented in genic and exonic regions (Figure 30). Nevertheless, some polyTEs do rise to high allele frequencies and many also show high levels of geographic differentiation consistent with what has been seen for SNPs [103]. This differentiation is precisely what makes them good markers for ancestry inference, particularly when considered as an ensemble, but it also suggests the possibility polyTE insertions may influence population specific phenotypes shaped by selection. Additional analysis on the effects of selection on TE polymorphisms, as well as the relationship between polymorphic TEs and potentially adaptive phenotypes, will be needed to test this assertion.

3.5 Conclusions

Polymorphic TE loci have long been used as markers in human population genetic studies, and they are known to provide a number of advantages for such studies. The selection of which polyTE loci to use for population genetic studies has been largely *ad hoc*, based on a combination of literature and database surveys together with empirical evaluation on the suitability of individual loci as markers that can discriminate between populations. With the recent release of a genome-wide set of 16,192 TE polymorphisms by the 1000 Genomes

Project [103, 128], genotyped across 2,504 individuals from 26 global populations, it is now possible to systematically evaluate the utility of polyTE loci for human population genetic and ancestry studies. We have leveraged these newly released data to conduct the first genome-scale analysis of polyTE genotypes for the study of human genetic ancestry. We show that the genetic diversity represented by TE polymorphisms reflects known patterns of human evolution, and define sub-sets of polyTE loci that can be used as ancestry informative markers. We provide ranked lists of the polyTE loci than be used by researchers in the community for future ancestry and admixture analyses.

3.6 Materials and Methods

3.6.1 Transposable element polymorphisms

Human polymorphic transposable element (polyTE) genotypes were taken from the Phase3 data release (November 2014) of the 1000 Genomes Project [103, 128] (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/). These genotypes consist of phased presence/absence patterns of polyTE insertions at specific human genome sites for individual genomes, and they are characterized from human genome reference sequence mapped next-generation sequence data via 1) discordant read mapping for short paired-end reads and/or 2) split read mapping for longer reads as previously described[59]. PolyTE allele frequencies are calculated as the number of present TE insertions (TEi) normalized by the total number of sites in the population ($2n$) : $TEi/2n$. The extent to which individual polyTE loci differentiate populations was computed using the fixation index F_{ST} with the Weir Cockerham method [133] implemented in VCFtools [134] and the δ parameter [129],

which is defined as the absolute value of the difference in the allele frequencies between populations for TE polymorphisms.

3.6.2 Ancestry analysis

PolyTE-based allele sharing distances were computed for all pairs of human genomes by counting the total number of polyTE presence/absence alleles that differ between two individuals across all genomic insertion sites. Allele sharing distances computed in this way were projected in two-dimensional space using multi-dimensional scaling (MDS) implemented in R. This was done for pairwise distances computed between individual genomes and for average allele sharing distances among populations. Population average allele sharing distances were used to reconstruct a neighbor-joining[135] phylogenetic tree using the program MEGA6 [136].

3.6.3 Admixture analysis

The program ADMIXTURE was used to infer the proportion of ancestry contributions from ancestral populations to modern admixed populations from the Americas (ACB, ASW, CLM, MXL, PEL, PUR) based on polyTE genotypes. The program was first run in supervised mode with three ancestral clusters: African, Asian and European. Asian ancestry is taken here as a rough surrogate for Native American admixture in American populations given the relatively close evolutionary relationship between East Asian and Native American populations and the lack of Native American samples in the 1000

Genomes Project. PolyTE genotypes were then imputed for Native American genomes from the Human Genome Diversity Project [124, 137], using the impute panel from the 1000 Genomes Project with the program IMPUTE2 [138], and ADMIXTURE was run in supervised mode with the three ancestral clusters: African, European and Native American. The ancestry contribution fractions for modern admixed populations from the Americas computed based on observed Asian polyTE genotypes and imputed Native American genotypes were correlated to check for consistency.

3.6.4 Ancestry and admixture prediction analyses

The program ADMIXTURE was used together with a cross-validation approach in order to predict the ancestry of individuals based on their polyTE genotypes. The cross-validation method relied on an 80%/20% split of the data, whereby 80% of individual polyTE genotypes were used to build a three-cluster ancestry model with ADMIXTURE. The remaining 20% of individual polyTE genotypes were then tested against this model to predict their ancestry membership in one of the three groups. Group-specific ancestry was only assigned if the probability of group membership was calculated as ≥90%. Accuracy is then defined as the number of correct ancestry predictions normalized by the total number of predictions made. Error is defined as the root-mean-square difference (*RMSD*) between the predicted and actual ancestry inference made with the complete data. *RMSD* values are reported as the average prediction error for all individuals. This process was done repeatedly across individual polyTE genotypes based on decreasing numbers of

polyTE sites, from 500 to 10 in steps of 10. For each polyTE set, this 80/20 prediction process was repeated 100 times.

An analogous prediction approach was used to infer the continental ancestry contributions to an admixed American population (ASW) using ADMIXTURE. In this case, the training was done using individual polyTE genotypes from ancestral populations (African and European) and the testing was done using polyTE genotypes from admixed ASW individuals. This was done first using all 16,192 polyTE loci and then for individual polyTE genotypes based on decreasing numbers of polyTE sites, from 500 to 10 in steps of 10. The predicted ancestry contributions to admixed individuals were compared for results based on all polyTE loci and results based on reduced sets of polyTE loci using the root-mean-square difference (*RMSD*) for the African and European fractional ancestry contributions.

CHAPTER 4. POPULATION-SPECIFIC POSITIVE SELECTION ON HUMAN TRANSPOSABLE ELEMENT INSERTIONS

4.1 Abstract

Insertional activity of transposable elements (TEs) has had a major impact on the human genome; more than two-thirds of the genome is derived from TE sequences. Several families of human TEs – primarily Alu, L1 and SVA – continue to actively transpose, thereby generating insertional polymorphisms within and between populations. We analyzed the population genetic variation caused by human TE activity in an effort to understand how natural selection acts on TE polymorphisms. Our genome-wide study of selection on human TE polymorphisms entailed the analysis of 14,384 insertions among 1,511 individuals from 15 populations. Consistent with previous reports, allele frequencies and patterns of TE insertion polymorphisms are largely consistent with the action of negative (purifying) selection. Nevertheless, application of a modified population branch statistic test uncovered a number of cases of where polymorphic TE insertions have increased in frequency, for specific continental population groups, owing to the effects of positive (adaptive) selection.

4.2 Introduction

One of the major findings from the Human Genome Project was the extent to which the genome sequence was found to be derived from transposable element (TE) insertions. Initial analysis of the genome draft sequence, using the RepeatMasker program, indicated that 47% of the human genome was derived from TEs [19]. Of course, this result did not

come as a surprise to TE researchers, but it did underscore the potential impact of TE-derived sequences on the evolution, structure and function of the human genome for the broader research community. Subsequent studies relying on more sensitive sequence analysis methods have revised the fraction of TE-derived sequences in the human genome upwards, with a current estimate as high as 69% [2].

The ubiquity and abundance of TE-derived sequences in eukaryotic genomes, such as our own, begs an explanation. For years, the selfish DNA theory was held as the gold-standard explanation for the genomic presence of TEs. The selfish DNA theory posits that TEs are genomic parasites that provide no benefit for their hosts and exist simply by virtue of their ability to out-replicate the genomes in which they reside [139, 140]. This idea is based on the fact that since TEs replicate when they transpose, and are also inherited vertically across generations, they have an inherently biased transmission rate compared to host genes that rely exclusively on vertical transmission for their propagation. It was even shown that TEs' replicative advantage meant that they could, in theory, persist and spread in the face of a selective cost to their host genome [141].

The selfish DNA theory for TEs is closely linked to the notion that TE sequences should be either neutral genetic elements or subject to purifying selection. Given the fact that human TE activity entails the insertion of rather large pieces of DNA, ranging from several hundred to almost ten-thousand base pairs, it is entirely reasonable to expect TE insertions to be deleterious. There is in fact abundant evidence from studies of disease that human TE insertions can be highly deleterious. Human TE insertions have been linked to a number of diseases including rare Mendelian diseases as well as more common chronic diseases such as cancer [101, 142-146].

The numerous studies reporting deleterious effects of TE insertions can be considered to be consistent with the selfish DNA theory, with respect to the notion that TEs are genomic parasites, and clearly point to a role for purifying selection in countering their unchecked spread. However, in the years since the publication of the draft human genome sequence, there have been many other studies that have demonstrated how formerly selfish human TE sequences have been exapted [71], or domesticated [72], to play a functional role for their hosts. For the most part, these studies have uncovered a role for TE-derived sequences in the regulation of human genes [73]. TE-derived sequences have been shown to contribute a wide variety of regulatory sequences, including promoters [74-76], enhancers [77-81], transcription terminators [82] and several classes of small RNAs [83-85]. Human TEs also influence various aspects of chromatin structure throughout the genome [19, 86-90].

It is important to note that all of the aforementioned studies on TE-derived regulatory sequences have dealt exclusively with relatively ancient TE insertions that are fixed among human populations. In other words, all known examples of specific human TE-derived regulatory sequences will be found at the same genomic locations in any individual person. While such fixed TE-derived regulatory sequences are certainly functionally relevant, by definition they will not be a source of genetic regulatory variation between individuals. The fact that TE-derived regulatory sequences correspond to relatively ancient fixed TEs is not at all surprisingly when you consider that the vast majority of human TE sequences, ~99.2% by our own rough calculation, correspond to ancient TE families that are no longer capable of transposition. However, very recent developments in genomics and bioinformatics are just beginning to enable systematic, genome-scale surveys of human

polymorphic TE (polyTEs) with insertion site locations that vary among individuals. The 1000 Genomes Project (1KGP) in particular has resulted in a collection of 16,192 polyTE genotypes characterized for 2,504 individuals from 26 global populations. Analysis of this data set has the potential to yield novel insights regarding the role of natural selection in shaping human TE genetic variation.

There is abundant evidence of adaptive evolution of polyTEs in *Drosophila* [91-95] along with studies that show the regulatory potential of polyTEs in mice [96]. However, at this time there is only tentative evidence to suggest that human polyTEs have been subject to positive (adaptive) selection [97]. We took advantage of the recently released 1KGP polyTE data in order to evaluate the role that natural selection has played in shaping this understudied, but potentially impactful, source of human genetic variation. In particular, we were interested to measure the effect of natural selection on human TE genetic variation along with the potential connection between polyTE selection and genome regulation. To do so, we performed a comparative analysis on the polyTE insertion allele frequencies within and between major human population groups (Figure 31). This allowed us to evaluate the effect of both negative (purifying) and positive (adaptive) selection on polyTE genetic variation. In the case of positive selection, we developed and applied a modified version of the population branch statistic (PBS) test, paired with coalescent simulation of polyTE allele frequencies, in order to detect cases of polyTE insertions that have been swept to high allele frequencies in specific human populations.

4.3 Results and Discussion

4.3.1 Characterization of human polymorphic transposable elements (polyTEs)

There are three main families of active human TEs that generate insertion polymorphisms among individual human genomes [3]: L1 [27, 28], Alu [29, 30] and SVA [31, 32]. L1 (Long Interspersed Element-1, or LINE1) are 6-8 kb long, autonomous, non-LTR (long terminal repeat) retrotransposons [33-35]. Alus and SVAs are non-autonomous, non-LTR retrotransposons that are retrotransposed in *trans* via the L1 transposition machinery [36, 37]. Alus are short interspersed elements (SINEs) that are ~300 bp long [38, 39], whereas SVA are composite elements made up of SINE, VNTR (Variable number tandem repeat) [40, 41] and Alu elements and can vary from 100-1500 bp in length [42]. The Phase 3 release of the 1000 Genomes Project (1KGP) includes polymorphic transposable (polyTE) genotype calls for these three active TE families from 2,504 individuals sampled across 26 populations world-wide [42, 99].

The insertion site locations of polyTEs in the 1KGP sample donors' genomes, along with their presence/absence genotypes, were characterized from next-generation sequence data by the 1KGP Structural Variation Group using the computational tool MELT. The program MELT works by screening for discordant read mappings for short paired-end reads and split read mapping for longer reads. MELT's performance was previously benchmarked by its developers using an experimentally validated set of polyTEs characterized for a single 1KGP individual, and the polyTE genotype calls from MELT were found to be quite reliable [42]. In addition, our own group independently benchmarked the performance of MELT and validated the accuracy of the human polyTE

genotype calls that it generates [147]. In our hands, MELT showed 90.4% precision and 81.5% recall and was the top performer among 21 polyTE detection programs that were evaluated.

Table 6 Human populations analyzed in this study.

Global populations are organized into three continental groups and the numbers of individuals analyzed for each population are shown. Population names and descriptions follow the conventions of the 1000 Genomes Project.

	Short	Full Description	*n*
African (*n*=504)	ESN	Esan in Nigeria	99
	GWD	Gambian in Western Division, The Gambia	113
	LWK	Luhya in Webuye, Kenya	99
	MSL	Mende in Sierra Leone	85
	YRI	Yoruba in Ibadan, Nigeria	108
Asian (*n*=504)	CDX	Chinese Dai in Xishuangbanna, China	93
	CHB	Han Chinese in Bejing, China	103
	CHS	Southern Han Chinese, China	105
	JPT	Japanese in Tokyo, Japan	104
	KHV	Kinh in Ho Chi Minh City, Vietnam	99
European (*n*=503)	CEU	Utah residents with Northern and Western European ancestry	99
	FIN	Finnish in Finland	99
	GBR	British in England and Scotland	91
	IBS	Iberian populations in Spain	107
	TSI	Toscani in Italy	107

The 26 populations from the 1KGP can be organized into 5 major continental population groups. The African, Asian and European continental population groups consist of (relatively) non-admixed individuals, and polyTE genotypes from these groups were analyzed here for the purpose of measuring selection on polyTEs (Figure 32). We analyzed a total of 14,384 polyTE genotypes from 1,511 individuals across 15 individual populations from these three continental population groups (Table 6). PolyTE genotype calls from the

three most actively transposing families of TEs were evaluated: Alu (11,216 or 78.0%), L1 (2,421 or 16.8%) and SVA (747 or 5.2%).

4.3.2 Negative selection on human polyTEs

PolyTE genotype calls were used to calculate insertion allele frequencies within and between populations in order to measure the effects of natural selection on human genetic variation caused by recent TE activity (see Materials and Methods). Consistent with the results of our previous study on human polyTEs [102], we found several lines of evidence in support of the action of negative (purifying) selection on human-specific TE insertions. These results are not surprising given the known deleterious effects of human TE insertions[101, 142-146], but they can also be considered to provide an additional line of support for the reliability of the polyTE genotype calls used to generate the allele frequencies analyzed here.

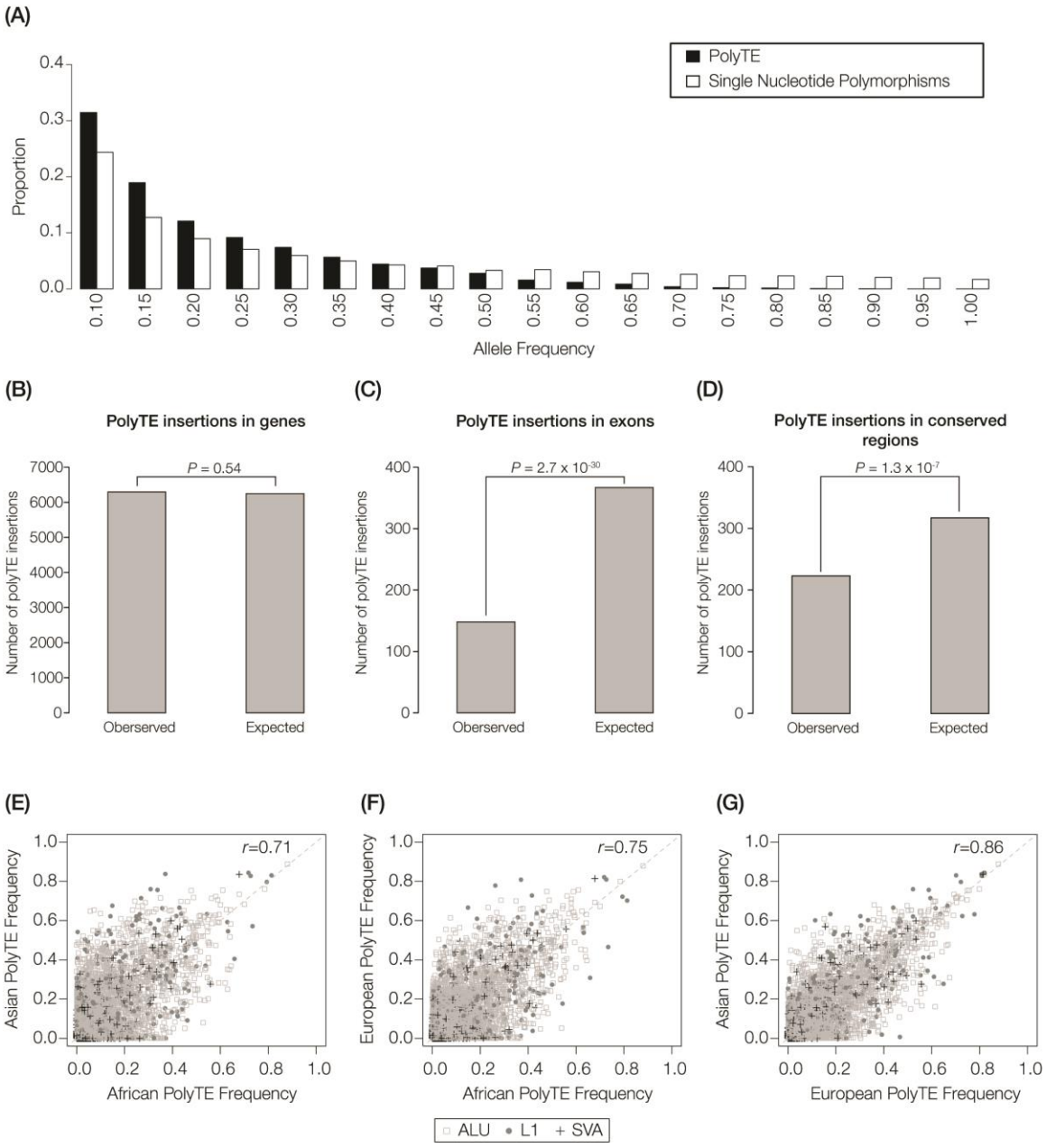

Figure 16 Signatures of purifying selection on polyTE insertions.

(A) Unfolded allele frequency spectrum for polyTE insertions (black bars) and SNPs (white bars). Observed versus expected counts of polyTE insertions in genes (B), exons (C) and conserved regions (D). The significance of the differences between observed versus expected TE counts (Fisher's exact test P-values) are shown for each plot. (E-G) Correlations of polyTE insertion allele frequencies between continental population groups are shown for shared Alu, L1 and SVA insertions; Spearman correlation coefficients are shown as r-values.

The majority of polyTE insertions show low allele frequencies; 11,658 (81.0%) polyTE loci exhibit average allele frequencies of less than 5% across all three continental population groups, and 10,119 (70.3%) exhibit allele frequencies less than 5% within each of the continental groups. Accordingly, polyTE insertions show a highly skewed allele frequency distribution with relatively fewer high frequency alleles compared with what can be seen for single nucleotide polymorphisms (SNPs) (Figure 16A). When the Alu, L1 and SVA polyTE families are considered separately, they all show similarly left skewed allele frequency distributions (Figure 33). Skewed allele frequency distributions of this kind are consistent with purifying selection acting to keep polyTE insertions at low frequencies.

PolyTE insertions also show evidence of being excluded from functionally important regions by purifying selection. While the observed number of polyTE insertions within genes is not statistically distinguishable from the expected number (Figure 16B), there is a highly significant deficit of polyTE insertions within exons compared to what is expected (Figure 16C). These results indicate that polyTE insertions are allowed to accumulate in functionally inert regions, such as introns, but removed from more functionally critical exonic regions by selection. There is a similar deficit of observed compared to expected polyTE insertions within evolutionary conserved regions (Figure 16C), which are also considered to be functionally important [148, 149].

The allele frequencies of polyTE insertions that are shared among continental population groups are both skewed towards low frequencies and highly correlated between groups (Figure 16 E-F and Figure 34). These results are consistent with both the action of purifying selection, to keep polyTE insertion allele frequencies low overall, and genetic

drift allowing less constrained insertions to increase in frequency at similar rates among populations.

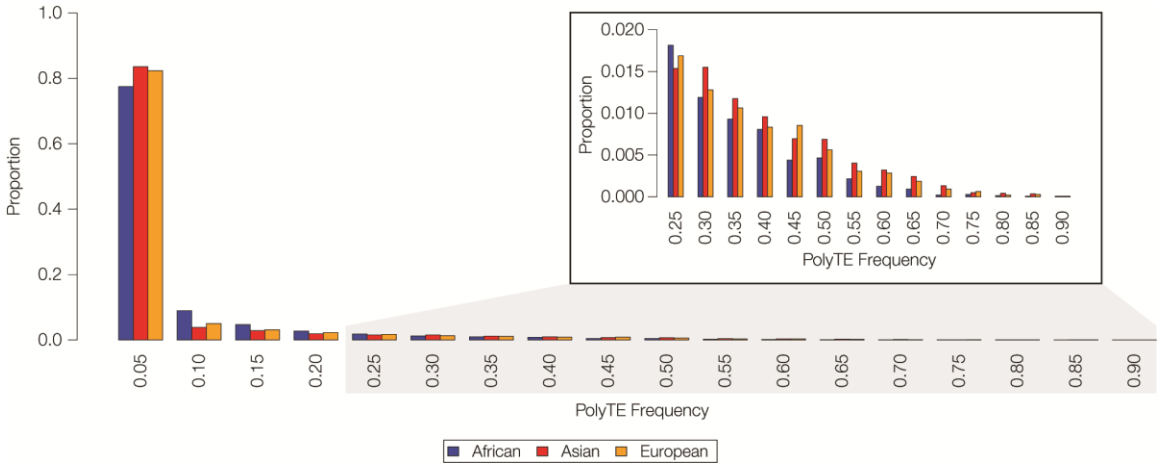

Figure 17 Unfolded allele frequency spectrum for polyTE insertions from African (blue), Asian (red) and European (gold) population groups.

The inset expands the higher range of the allele frequency spectrum (≥0.25).

4.3.3 Detecting positive selection on human polyTEs

Decomposition of the polyTE allele frequency spectrum into continental population group-specific spectra revealed an unexpected pattern that suggested the possibility that some human polyTE insertions may have increased in frequency owing to the effects of positive selection. The population group-specific polyTE allele frequency spectra are also highly skewed to the low end of the distribution; however, there are a number of polyTE insertions, particularly in Asian and European populations, that appear to be found at higher than expected frequencies (Figure 17). At the high end of the polyTE allele frequency spectrum, there is a shift whereby Asian and European polyTEs become relatively more

frequent than African TEs. This same shift is not seen for the allele frequency distribution of intergenic SNPs, which can be considered as a surrogate of neutral evolution (Figure 35). The shift to a greater proportion of Asian and European polyTEs at the high end of the allele frequency spectrum is unexpected given the fact that African populations are ancestral, and thus their polyTEs should have had more time to drift to higher frequencies. Nevertheless, this pattern could be attributable to less efficacious selection in Asian and European populations due to historically lower effective population sizes in these groups [150].

We developed and applied a modified version of the population branch statistic (PBS) test in order to try and distinguish between neutral evolution of polyTE insertions (*i.e.* genetic drift) versus population group-specific increases in polyTE allele frequencies that can be attributed to positive selection. We chose this method given its demonstrated power to detect recent positive selection of SNPs in human populations [151]. The PBS test measures population-specific divergence levels by converting pairwise F_{ST} values into population-specific branch lengths, and we adopted this method by computing the F_{ST} values from polyTE allele frequencies as described in the Materials and Methods. Deviations from neutrality are detected as extreme population-specific branch length values using this approach.

Figure 18A shows the genome-wide polyTE PBS tree with average branch lengths for each continental population group. On average, shared polyTE insertions show higher PBS values in Africa, compared to Asia and Europe, consistent with the fact that African populations are an outgroup to the more recently diverged Asian and European populations. PBS branch length distributions for each continental population group are highly skewed;

the vast majority of polyTEs have low PBS values for all three populations (Figure 18B). These results are consistent with the low overall allele frequencies observed for polyTEs (Figure 16A and Figure 33); in other words, the majority of polyTE PBS trees do not appear to show evidence for positive selection.

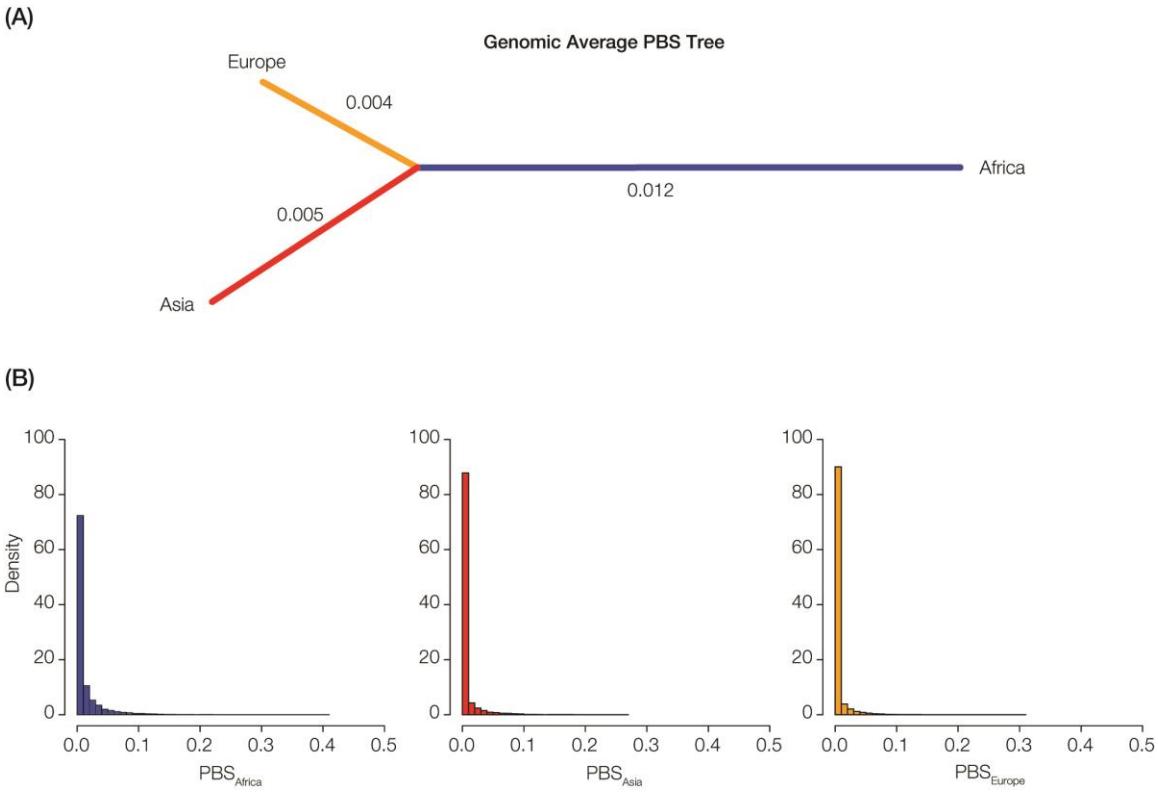

Figure 18 Overview of the population branch statistic (PBS) test metric used to detect positive selection on polyTE insertions.

(A) Tree constructed with branch lengths from the genomic averages of continental group-specific PBS values. (B) Histograms showing the PBS branch length distributions for the African (blue), Asian (red) and European (gold) population groups.

We further evaluated all of the PBS trees in an effort to look for rare cases of positively selected polyTEs. To do this, the observed population group-specific branch lengths from

the polyTE PBS trees were compared to a null distribution of branch lengths generated via coalescent model simulations. A coalescent model – consisting of a tree for the three continental population groups along with estimated divergence times and effective population sizes – was used to simulated polyTE allele frequencies as described in the Materials and Methods (Figure 19 A-B). The parameter values for the human population coalescent model that we used were taken from a recent study that estimated these parameters using a large set of neutrally evolving loci along with statistically rigorous approaches [150]; as such, these parameter values are considered to be a good approximation for human population divergence. The coalescent simulated polyTE allele frequencies were in turned used to create individual polyTE PBS tree branch lengths. The set of simulated PBS branch lengths was then compared to the observed set in order to look for statistically significant outliers that represent putative positively selected polyTEs. The coalescent approach for generating a null distribution was chosen in an effort to minimize the possibility of observing extreme population-specific PBS branch lengths that are nevertheless consistent genetic drift.

The coalescent simulation generates ancestral and extant polyTE allele frequencies that are highly correlated (Figure 19C), a result that is consistent with the observed polyTE frequencies (Figure 16E-G and Figure 34). In addition, we observe that coalescent simulations starting from the same ancestral polyTE insertion allele frequencies can generate very different extant allele frequencies (Figure 19D). Both of these results underscore the conservative nature of the coalescent approach we used to generate a null distribution of PBS branch lengths. Statistical comparison of the observed versus simulated sets of PBS branch lengths yielded a set of 163 polyTE insertions (1.13% of the

full set) that appears to have increased in frequency in the Asian or European population groups, based on positive selection. Among these 163 putatively selected polyTE insertions, only 79 have frequencies > 10% and only 14 have frequencies > 25% in either of the Asian or European populations. Furthermore, the entire set of 163 putatively selected polyTE insertions is not enriched (or depleted) for any particular TE family type, any genomic region or any particular class of functionally important (regulatory) genomic elements.

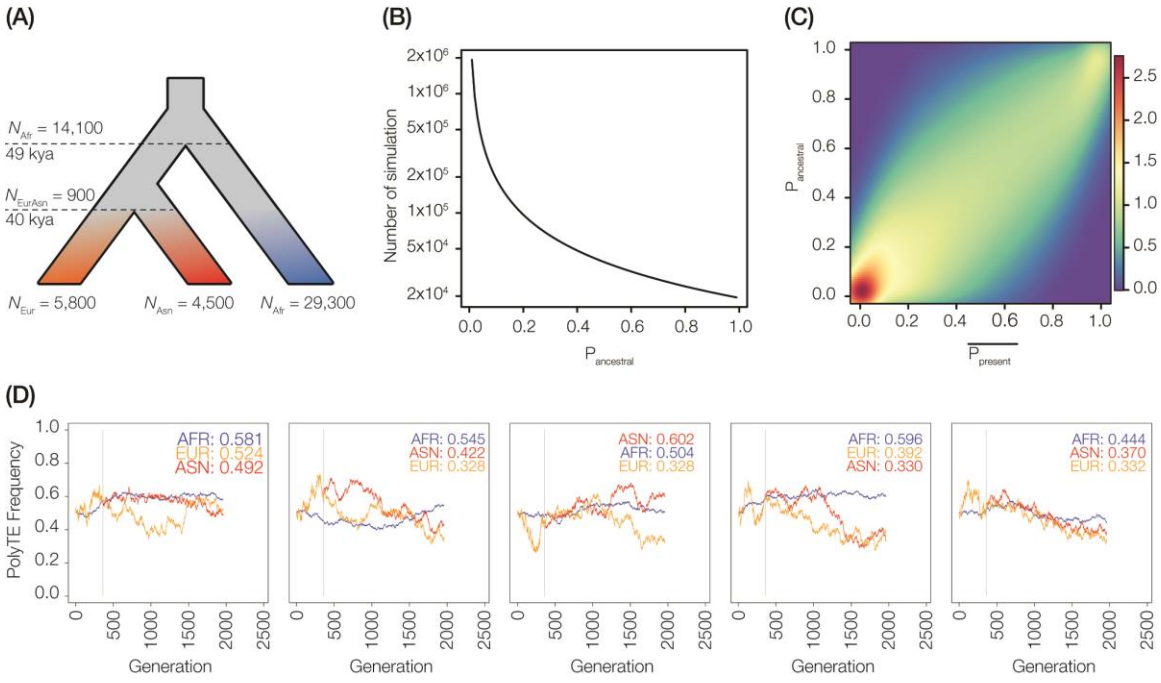

Figure 19 Coalescent modelling of polyTE insertion allele frequencies.

Coalescent modelling was used to generate a null distribution of PBS values for the purpose of detecting positive selection on polyTE insertions. (A) Scheme of the coalescent model and parameters used to simulate polyTE insertion allele frequencies. The coalescent model consists of the tree shown, the effective population sizes (N) at each node of the tree and the number of thousands of years ago (kya) that correspond to the two population splits in the tree: (i) out of Africa and (ii) Europe-Asia. (B) The number of coalescent model simulations run (y-axis) is plotted for each initial ancestral polyTE insertion frequency ($P_{ancestral}$ on the x-axis), ranging from 0.01 to 0.99. (C) Density scatter plot comparing the coalescent model ancestral polyTE insertion frequencies ($P_{ancestral}$ on the y-

axis) to the mean of the extant polyTE insertion frequencies ($P_{present}$ on the x-axis) for the three simulated population groups. (D) Five examples of coalescent models run with initial polyTE frequencies of 0.5. The plots show the polyTE insertion frequency dynamics across generations for each coalescent model run. The final (extant) polyTE frequency values of each coalescent model run are shown for each population group: African (AFR-blue), Asia (ASN-red) and European (EUR-gold).

Given the somewhat ambiguous results regarding all of polyTEs with significant PBS test scores, it is difficult to evaluate the robustness of the findings for most of the putatively selected polyTEs, and it is not possible to infer any potential functional role that many of these insertions may play in the genome. Accordingly, we chose to focus on a limited set of putatively selected polyTEs, for which multiple lines of evidence support both the action of positive selection on the insertions as well as some potential functional (regulatory) significance. To implement such a composite approach, we searched for putatively selected polyTE loci that were found at anomalously high frequencies within a single population group and were also co-located within genes and/or functionally important genomic regulatory elements. A list of seven positively selected polyTEs that fit these criteria are shown in Table 7, and a number of examples from this table are described further in the next section.

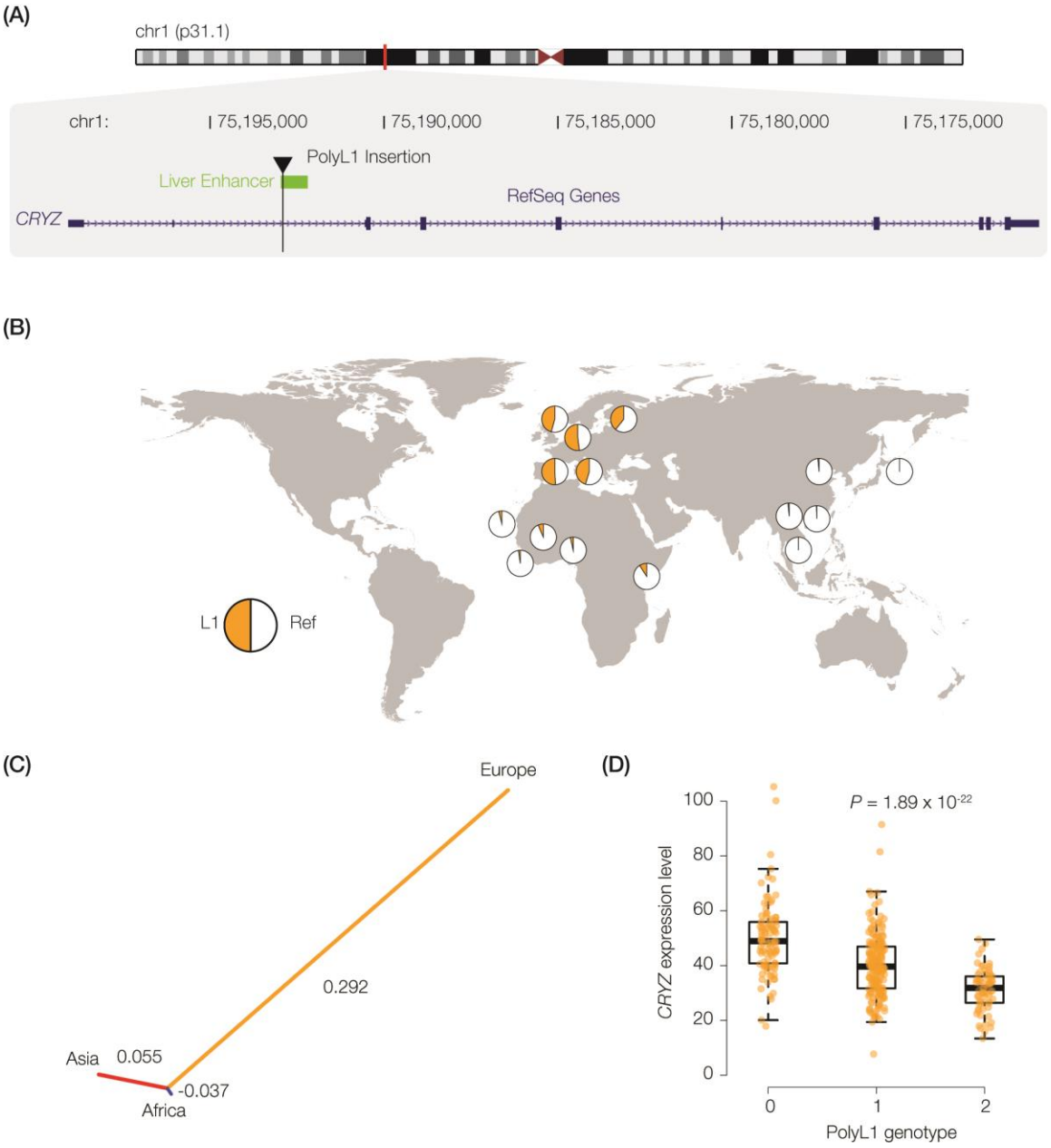

Figure 20 Positively selected polyL1 insertion in the CRYZ gene.

(A) Chromosome 1 ideogram showing the location (red bar) of the CRYZ gene on the short arm of chromosome 1 along with a CRYZ gene model showing the location of the polyL1 insertion and its co-located liver enhancer element (green bar). (B) Frequencies of the European selected polyL1 insertion (gold in the pie charts) for the individual populations studied here from Africa, Asia and Europe. (C) Tree with branch lengths scaled to the population group-specific PBS values (shown for each branch). (D) CRYZ expression level distributions are shown for European individuals that have 0, 1 or 2 copies of the selected polyL1 insertion. The significance of the differences in expression among individuals for

the three different polyL1 insertion genotypes between is shown (linear additive model P-value).

4.3.4 Examples of positively selected human polyTEs

One of the most promising candidates for positive selection is a polymorphic L1 (polyL1) insertion located on the short arm of chr1 at position 75,192,907 (Figure 20A). This polyL1 is inserted within the second intron of the crystallin zeta gene (*CRYZ*, also known as Quinone Reductase or *QR*) and co-located with a liver enhancer element [90]. This polyL1 is an ancient insertion that is found in all 26 populations sampled as part of the 1KGP; however, it is found low frequencies in the African (5%) and Asian (1%) population groups. There was a striking increase in the allele frequency of this insertion along the European lineage, and it is currently found at an average allele of 47% in European populations (Figure 19B). When these polyTE allele frequencies are used to calculate the F_{ST} values that underlie the PBS test, the European-specific branch on the PBS tree is extremely long compared with the African and Asian branches (Figure 19C). Comparison of this observed polyL1 PBS tree to the set of coalescent simulated tress, with similar average branch lengths, yields an FDR q-value of 0.019 (Table 7). Consistent with the potential regulatory effects of this insertion, expression quantitative trait loci (eQTL) analysis shows that the presence of this specific polyL1 insertion in European individuals is significantly associated with lower expression of the *CRYZ* gene in B-lymphoblastoid cell lines (Figure 19D).

Another strong candidate for positive selection is a polymorphic Alu (polyAlu) insertion at chr16 position 75,655,176 (Figure 36A). This polyAlu is located within the second

intron of the Adenosine Deaminase, tRNA Specific 1 gene (*ADAT1*) and is co-located with enhancer elements predicted to have activity in numerous cell lines analyzed by The Roadmap Epigenomics project [90]. This polyAlu insertion is also an ancient insertion that is found in all human populations surveyed by the 1KGP. It is seen at low frequencies in African (4%) and European (5%) population groups and far higher frequency in the Asian population group (44%) (Figure 36B). Accordingly, the Asian-specific branch on the PBS tree is far longer than the African or European branches (Figure 36C), and comparison with coalescent simulated trees yields an FDR q-value of 0.04. This is a clear case of a marked increase in polyTE allele frequency that cannot be readily explained by genetic drift. In addition, the location of the insertion is suggestive of regulatory function; although, a lack of gene expression data from matched Asian samples does not allow us to directly assess the association of the insertion with changes in *ADAT1* expression.

Table 7 List of high confidence positively selected polyTEs.

Chr[1]	Position[2]	Family[3]	f_{AFR}[4]	f_{ASN}[5]	f_{EUR}[6]	PBS[7]	q-value[8]	Cont[9]	Gene[10]	Enh[11]	TFBS[12]	eQTL[13]
1	75,192,907	L1	0.05	0.01	0.47	0.29	0.019	EUR	CRYZ	Yes	.	CRYZ
1	169,442,974	ALU	0.03	0.35	0.02	0.19	0.046	ASN	SLC19A2	Yes	.	.
4	43,399,986	ALU	0.08	0.16	0.61	0.31	0.033	EUR	.	.	.	.
11	10,042,452	L1	0.01	0.16	0.01	0.08	0.039	ASN	SBF2	Yes	.	.
14	88,415,499	L1	0.02	0.10	0.00	0.05	0.033	ASN	GALC	Yes	.	GPR65
16	75,655,176	ALU	0.05	0.44	0.04	0.24	0.040	ASN	ADAT1	Yes	Yes	.
17	44,153,977	SVA	0.00	0.00	0.21	0.13	0.031	EUR	KANSL1	.	.	KANSL1

[1]Chromosome
[2]Base position in the hg19 human genome reference assembly
[3]PolyTE family
[4,5,6]PolyTE insertion frequency in the African, Asian and European population groups
[7]FDR corrected q-value for PBS selection test based on the coalescent simulation (Figure 35)
[8]Continental population group in which the polyTE is selected
[9]Gene name in which the selected polyTE insertion is located
[10]Selected polyTE insertion located in an enhancer
[11]Selected polyTE insertion located in a transcription factor binding site (TFBS)
[12]Target gene name for which the selected polyTE insertion is an eQTL

The candidate with the strongest PBS-based evidence of positive selection is a polyAlu insertion at chr4 position 43,399,986 (Figure 37A). This Alu element is inserted in an intergenic region and does not overlap with any known functional (regulatory) elements. Nevertheless, its relative allele frequencies leave little doubt as to the role for positive selection in shaping its population-specific patterns of variation. This polyAlu insertion is found in all of the continental groups with an allele frequency of at least 5% in each of the 26 1KGP populations. It shows the highest population-specific frequency for any of the putatively selected insertions, with 61% average frequency in the European populations compared to 8% and 16% in the African and Asian population groups, respectively.

Accordingly, it also has the highest PBS test statistic value for any of the high confidence positively selected polyTEs shown in Table 7.

In addition to the three examples described above, putatively selected polyTE insertions in four other genes – *SLC19A2*, *SBF2*, *GALC* and *KANSL1* – also showed strong composite signals of positive selection (Table 7). Putatively selected polyTE insertions in the first three genes (*SLC19A2*, *SBF2* and *GALC*) overlap with enhancer elements while insertions in the last two genes (*GALC* and *KANSL1*) were found to behave as eQTLs to *GPR65* and *KANSL1* genes, respectively.

4.4 Materials and Methods

4.4.1 Polymorphic transposable element (polyTE) analysis

Genotype calls for 14,384 human polymorphic transposable element (polyTE) insertions were obtained from the Phase 3 data release of the 1000 Genomes Project (1KGP) [42, 99, 116], with locations corresponding to build GRCh37/hg19 of the human genome reference sequence. PolyTE genotype calls report the presence or absence of insertions for members of three families of human polyTEs: Alu, L1 and SVA. For any given polyTE insertion site, individuals can be homozygous absent (0 insertions), heterozygous (1 insertion) or homozygous present (2 insertions). PolyTE genotype calls were taken for 1,511 individuals from 15 populations corresponding to 3 non-admixed continental population groups: Africa, Asia and Europe (Table 6 and Figure 32). For each polyTE insertion site, its polyTE allele frequency was calculated as the total number of TE insertions observed

at that site (TE_i) normalized by the total number of chromosomes in the population under consideration ($2n$): $TE_i/2n$. PolyTE insertion site allele frequencies were calculated separately for all 15 individual population groups as well as for the 3 continental population groups.

The BEDTools [152] program was used to compare the locations of polyTE insertions to (1) the genomic coordinates of RefSeq genes [153] (transcription start to transcription stop site for each gene), (2) the locations of RefSeq gene exons, and (3) the locations of conserved genomic regions. Conserved genomic regions were characterized using GERP++ RS conservation scores [149] taken from the UCSC Genome Browser [154], with GERP++ RS > 3 taken to represent conserved genomic regions. The observed counts of polyTE insertions for each of these three functional features – genes, exons and conserved regions – were compared to the expected counts, which were computed as the total number of polyTE insertions multiplied by the fraction of the genome occupied by each feature. The significance of the differences in the observed versus expected counts of polyTE insertions for each feature were calculated using Fisher's exact test. All statistical analyses and correlations were performed in R.

4.4.2 Population branch statistic (PBS) calculation

For each polyTE insertion, its continental population group-specific allele frequencies were used to calculate African, Asian and European population branch statistic (PBS) values. PBS values were calculated based on pairwise polyTE frequency F-statistics (F_{ST}) [151] as shown:

$$F_{ST} = \frac{(H_T - H_S)}{H_T} \quad (2)$$

$$T = -\log(1 - F_{ST}) \quad (3)$$

$$PBS_{Africa} = \frac{(T^{AS} + T^{AE} - T^{SE})}{2};\ PBS_{Asia} = \frac{(T^{AS} + T^{SE} - T^{AE})}{2}\ ;\ PBS_{Europe} = \frac{(T^{SE} + T^{AE} - T^{AS})}{2} \quad (4)$$

where,

H_S is the sample polyTE heterozygosity within each continental population group being compared

H_T is the total polyTE heterozygosity for both continental population groups being compared

T^{XY} is the polyTE divergence level for continental population groups X and Y being compared

T^{AS}, T^{AE} and T^{SE} denote the polyTE divergence levels between all three pairs of continental groups compared: Africa-Asia (AS), Africa-Europe (AE) and Asia-Europe (SE)

4.4.3 Detection of positively selected polyTEs using PBS values and coalescent modelling

Observed polyTE insertion PBS values were compared to a null distribution of values generated via coalescent modelling in order to detect positively selected polyTEs. A Wright-Fisher based human coalescent model with two population divergence events, yielding the three extant continental population groups analyzed here, was implemented for this purpose (Figure 19). Model parameter values for – (1) the time elapsed since the population divergence events and (2) the effective population sizes – were taken from a previous report by Gronau, et al [150]. The coalescent model was used to simulate polyTE insertion frequency dynamics starting with ancestral frequencies (p) ranging from 0.01 to 0.99, incrementing by steps of 0.01. The number of simulations (s_i) for each ancestral frequency (p_i) was performed proportional to $1/p_i$ such that a total number of 10 million simulated instances of continental population group-specific extant polyTE frequencies were generated:

$$s_i = int\left(\frac{10{,}000{,}000}{p_i \times \sum_{0.01}^{0.99} p}\right) \quad (1)$$

PolyTE frequencies simulated in this way were then used to calculate simulated PBS values, in the same way as described in the previous section, and the simulated PBS values were used to form a null distribution for statistical testing. For the purposes of statistical testing, simulated and observed PBS trees with similar mean branch lengths were compared and the deviation of the observed versus simulated continental population group-specific branch lengths were calculated. Since this procedure entailed multiple statistical tests, false discovery rate q-values were used to establish statistical significance.

4.4.4 Gene regulatory potential of selected polyTEs

The locations of polyTE insertions that show evidence for positive selection were compared to several classes of gene regulatory features and functional genomic data. Computationally inferred enhancer locations from 125 cell lines were obtained from The Roadmap Epigenomics Project [90], and transcription factor binding site locations were obtained from the UCSC Genome Browser Txn Factor ChIP track. The locations of enhancer elements were computationally inferred using the core 15-state model from five chromatin marks assayed for 128 epigenomes across 30 different cell types [90]. Human gene expression levels for 358 individuals from four European 1KGP populations (CEU, FIN, TSI, GBR) were obtained from the RNA-seq analysis performed by the GUEDVADIS project [155, 156], [157]. Individuals' polyTE genotypes were compared to their gene expression levels to identify expression quantitative trait loci (eQTL) that correspond to polyTE insertion sites using the program Matrix eQTL [158]. Matrix eQTL was run using the additive linear (least squares) model with covariates for gender and population.

4.5 Conclusions

We present here the first systematic, genome-wide study of the effects of natural selection on human genetic variation that results from the recent activity of TEs. The majority of human polyTE insertions are found at low allele frequencies, within and between populations, and appear to evolve via negative (purifying) selection, with others increasing to moderate allele frequencies via genetic drift. Nevertheless, a small, but not insubstantial,

minority of polyTE insertions show patterns of allele frequencies that are consistent with population-specific positive selection. A number of these positively selected TEs have functional features that are consistent with a role in human gene regulation. These results indicate that the exaptation of human TE sequences, which was previously limited to relatively ancient and fixed TE sequences, can also occur for more recently active polyTEs with insertion sites that vary among individuals within and between populations.

CHAPTER 5. POPULATION AND CLINICAL GENETICS OF HUMAN TRANSPOSABLE ELEMENTS IN THE (POST) GENOMIC ERA

5.1 Abstract

Recent technological developments – in genomics, bioinformatics and high-throughput experimental techniques – are providing opportunities to study ongoing human transposable element (TE) activity at an unprecedented level of detail. It is now possible to characterize genome-wide collections of TE insertion sites for multiple human individuals, within and between populations, and for a variety of tissue types. Comparison of TE insertion site profiles between individuals captures the germline activity of TEs and reveals insertion site variants that segregate as polymorphisms among human populations, whereas comparison among tissue types ascertains somatic TE activity that generates cellular heterogeneity. In this review, we provide an overview of these new technologies and explore their implications for population and clinical genetic studies of human TEs. We cover both recent published results on human TE insertion activity as well as the prospects for future TE studies related to human evolution and health.

5.2 Human transposable element research in the (post) genomic era

5.2.1 Technology driven research and discovery on human transposable elements

A convergence of new technologies in three key areas – genomics, bioinformatics and high-throughput experimental techniques – is providing unprecedented opportunities for research and discovery on population and clinical genetic aspects of human transposable elements (TEs). In this review, we briefly cover these exciting technological developments and explore their implications for understanding how the activity of human TEs impacts the evolution and health of the global population. We would like to emphasize that our treatment is by no means intended as an exhaustive review of the subject, rather we are simply attempting to call the readers' attention to what we perceive to be some of the most relevant developments in this area along with the potential for future studies that these advances entail. It should also be noted that the review is focused primarily on the new bioinformatics tools that can be used to detect polymorphic TE insertions from next-generation sequence data, rather than the high-throughput experimental techniques, since we are most familiar with the computational approaches.

Developments in genomics technology, and next-generation sequencing in particular, have taken us from the analysis of a single human genome, which alone has provided profound insight into the biology of human TEs, to the population genomics era where whole genome sequences from thousands of human individuals can be compared. Concomitant developments of bioinformatics tools for genome sequence analysis have allowed for the discovery and characterization of the genetic variants that are generated via recent TE activity, *i.e.* human TE polymorphisms, via the comparative analysis of next-generation re-

sequencing data from multiple human genomes. Finally, a suite of novel high-throughput experimental techniques, which also leverage next-generation sequencing data, have been developed and applied for the characterization of human polymorphic TE insertions at the scale of whole genomes across numerous samples.

The initial analysis of the first draft of the human genome sequence was, in some sense, a watershed event for TE research. One of the most significant findings of this research was the large fraction of the human genome that was shown to be derived from TE sequences; 47% of the genome sequence was reported to be TE-derived with a single family of elements, LINE-1 (L1), making up ~17% of the genome and another family, Alu, contributing almost 11 million individual copies[19]. These remarkable results were generated using homology-based sequence analysis with the program RepeatMasker[107]. Subsequent analysis of the human genome sequence, using a more sensitive *ab initio* algorithmic approach, has revised the estimate upwards to more than two-thirds of genome being characterized as TE-derived[2]. The abundance of TE sequences found in the human genome almost surely did not come as a surprise to members of the TE research community, but this finding certainly did underscore the potentially far reaching impact of these often underappreciated genetic elements on the human condition.

The 1000 Genomes Project (1KGP) can be considered as the successor to the initial human genome project as well as the initiative that ushered human genomic research into the so-called post genomics era[99, 103, 128]. As its name implies, the 1KGP entailed the characterization of whole genome sequences from numerous human individuals, and it did so with an eye towards capturing a broad swath of world-wide human genome sequence diversity. The 1KGP resulted in the characterization of whole genome sequences for 2,504

individual donors sampled from 26 global populations, which can be organized into 5 major continental population groups. The project was executed in three phases, each of which included a substantial focus on technology development, not only with respect to sequencing methods but also for the computational techniques that are needed to call sequence variants from next-generation re-sequencing data. This focus on technology development ultimately led to the characterization of genome-wide collections of human polymorphic TE (polyTE) insertion genotypes for all individuals in the project[42, 59]. Importantly, these data have been released into the public domain, thereby facilitating population and clinical genetic studies of human TE polymorphisms.

Advances in next-generation sequencing technology have also facilitated the development of high-throughput experimental techniques that can be used to detect *de novo* TE insertions, genome-wide across multiple samples. These high-throughput experimental techniques couple enrichment for sequences that are unique to active families of human TEs with subsequent next-generation sequencing and mapping techniques in order to discover the locations of novel TE insertions. Notably, these innovative experimental approaches have been successfully applied towards the characterization of somatic human TE activity in a variety of tissues, along with its potential role in cancer, as is discussed later in this review.

5.3 Active families of human TEs

As described above, a large fraction of the human genome sequence has been derived from millions of individual TE insertions. The process of TE insertion and accumulation in the

genome has taken place over many millions of years along the evolutionary lineage that led to modern humans, and it turns out that the vast majority of human TE-derived sequences were generated via relatively ancient insertion events. Most ancient TE insertions have accumulated numerous mutations since the time that they inserted in the genome, and as a consequence they are no longer capable of transposition. The vast majority of TE-derived sequences in the human genome (>99%) correspond to such formerly mobile elements. The most salient aspect of these inert human TEs, with respect to population and clinical genomics, is that their insertion locations are fixed in the human genome. In other words, each individual TE sequence insertion of this kind is found at the exact same genomic location in all human individuals and for all human populations. Thus by definition, these ancient and fixed TE sequences do not contribute to human genetic variation via insertion polymorphisms.

There are, however, several families of TEs that are still active in the human genome. Elements of the HERV-K, L1, Alu and SVA families remain capable of transposition and can thereby generate insertion polymorphisms among individual human genomes. The resulting TE insertion polymorphisms have important implications for human evolution and health (disease) as detailed later in this review. HERV-K and L1 are autonomous TEs that encode all of the enzymatic machinery needed to catalyze their own transposition, whereas Alu and SVA are non-autonomous elements that are transposed in *trans* by L1 encoded proteins[31, 36]. All four active families of human TEs correspond to retrotransposons that transpose via the reverse transcription of an RNA intermediate.

Members of the HERV-K family of active human TEs are human endogenous retroviruses, which are thought to have evolved from ancient retroviral infections that made their way

into the germline and eventually lost the capacity for inter-cellular infectivity via loss of coding capacity for the envelope protein. As such, HERV-K elements have genomic structures that are very similar to retroviruses, including long terminal repeat (LTR) sequences that flank the *gag* and *pol* open reading frames, which encode structural and enzymatic (integrase and reverse transcriptase) element proteins. L1 elements are long interspersed nuclear elements (LINEs) that are classified as non-LTR containing retrotransposons. Alu and SVA elements are both classified as short interspersed nuclear elements (SINEs). Alu elements are derived from 7SL RNA and are ~300 bp in length[38, 39]. SVAs are hybrid elements that are made up of SINE, VNTR (variable number tandem repeat) and Alu sequences and can vary from 100-1,500 bp in length[32, 40, 41].

5.4 Genome-scale characterization of TE insertions

5.4.1 Human genome sequencing initiatives

The initial draft of the human genome sequence took more than 10 years to complete at a cost of ~2.7 billion dollars[159]. Characterization of the human genome sequence was done with Sanger sequencing technology, using essentially the same chain termination biochemistry that was invented in the mid-1970s[160], albeit with refinements in automation. In the mid-2000s, staring with the Roche 454 pyrosequencing method, there was explosion of novel biochemical methods for DNA sequencing[161]. These so-called next-generation sequencing technologies enabled far higher throughput sequencing, at much lower cost, than the Sanger sequencing method used for the original human genome project. It is now possible to sequence an entire human genome in a single day at a cost of

~1,000 dollars using Illumina's patented sequencing by synthesis (SBS) technology. This hyper-exponential increase in sequencing capacity, and simultaneous decrease in its cost, is powering a series of human genome sequencing initiatives that have profound implications for the study of human TE genetic variation (Table 8).

The previously discussed 1KGP is the emblematic initiative for the characterization of whole human genome sequences at the population level; as such, it is difficult to overstate the impact that this project has had, and continues to have, on human population and clinical genomics. The 1KGP had the critical effect of stimulating experimental methods related to sequencing as well as numerous bioinformatics methods that are used for the analysis of genome sequence data, particularly as they relate to characterizing genetic variants. A major part of this effort was the development and refinement of methods for calling structural variants, including but not limited to TE insertion polymorphisms. Nevertheless, the 1KGP, which entailed the characterization of just over 2,000 whole genome sequences, has been dwarfed in scale by a number of subsequent initiatives that are currently underway (Table 8).

Table 8 Large scale genome sequencing initiatives.

Projects are sorted in descending order by the number of participants.

Project Name	PMID	# Participants	Description
Million Veteran Program (MVP)	26441289	1,000,000	Planned sequencing of 1 million U.S. Veterans (genotyping, whole genome and exome); current enrollment at 500k
SHGP	26583887	100,000	Catalogue of whole genome sequences of 100k Saudis
TOPMed	N/A	62,000	Sequencing of 62k individual genomes along with a variety of data for precision medicine initiative
UK10K	26367797	10,000	Sequencing of ~10k individuals from UK to inspect the effect of rare and low-frequency variants to human traits
Human Longevity	27702888	10,000	Deep sequencing of 10k human genomes; Data donated to Precision FDA
Iceland Genome Project	25807286	2,636	Catalogue of whole genome sequences of 2,636 Icelanders
1000 Genomes Project	26432245	2,504	International whole genome project that sampled 2,504 healthy individuals from 26 populations
EGDP	27654910	483	Catalogue of whole genome sequences of 483 genomes from 148 diverse population
SGDP	27654912	300	Catalogue of whole genome sequences of 300 genomes from 142 diverse population
GoNL	24974849	250	Catalogue of whole genome sequences of 250 Dutch parent-offspring families
Australian Aboriginals	27654914	108	Catalogue of whole genome sequences of 108 Aboriginal Australians

Several of the most ambitious human genome sequencing initiatives involve the characterization of cancer genome sequences. For example, the International Cancer Genome Consortium (ICGC) is collaborating with the US National Cancer Institute's The Cancer Genome Atlas (TCGA) to sequence genomes for 500 pairs of matched normal and tumor samples for 500 different tumor types, for an expected yield of 50,000 whole genome sequences[162, 163]. The US National Heart, Lung, and Blood Institute's (NHLBI) TOPMed precision medicine initiative is another health-related project that aims to sequence the genomes of 62,000 individuals[104]. There are a number of other large-scale human genome sequencing initiatives that are aimed at the populations of specific countries or global sets of populations. For example, the Wellcome Trust is sponsoring the UK10K initiative to sequence the genomes of 10,000 citizens of the United Kingdom[164], and Saudi Arabia intends to sequence 100,000 Saudi individuals for their own project[165]. The Simons Genome Diversity Project recently completed sequencing of 300 human genomes from 142 diverse populations[166], and the Estonian Genome Diversity Project sequenced 483 genomes from 148 populations[167]. Together, these projects, along with others like them, will provide a wealth of raw sequence data that can be mined for TE insertion polymorphisms using the computational and experimental approaches described in the sections that follow.

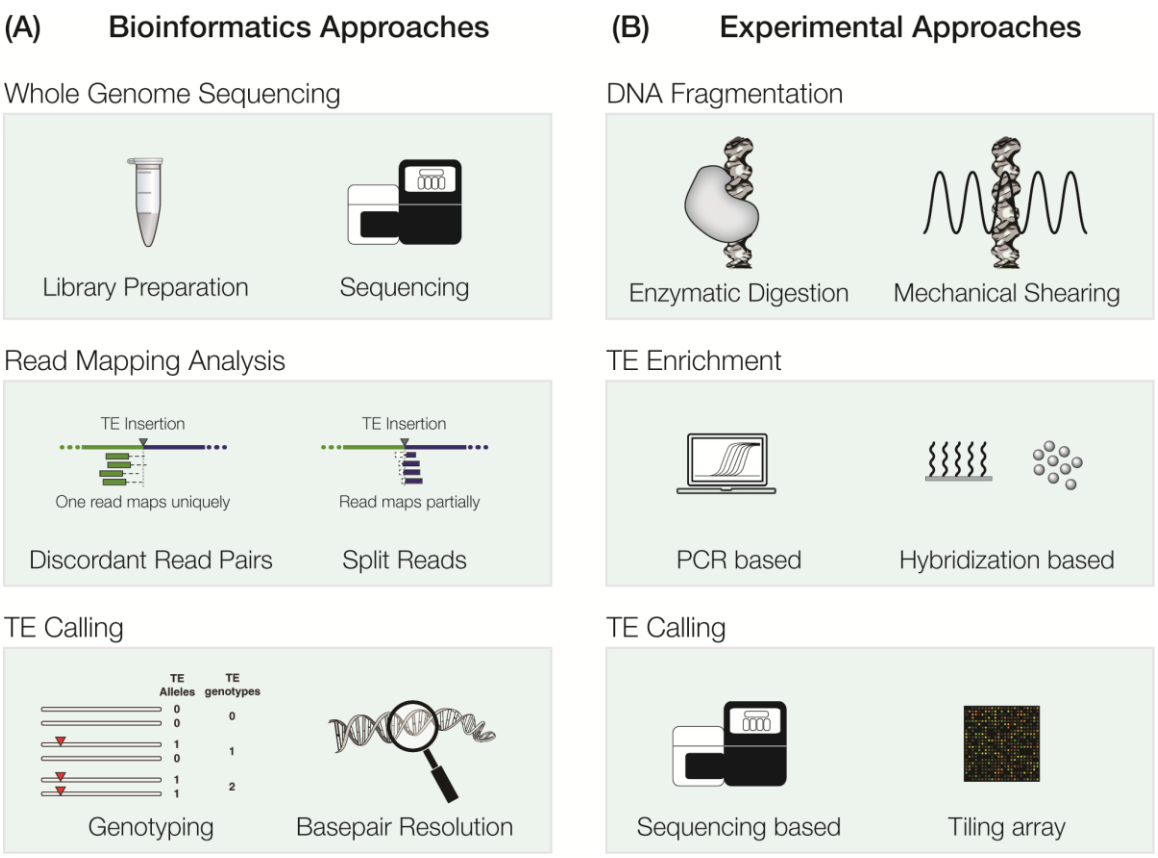

Figure 21 Schematic of the high-throughput bioinformatics (A) and experimental (B) approaches to human TE insertion discovery.

5.4.2 High-throughput techniques for TE insertion detection

5.4.2.1 Bioinformatics approaches

The characterization of single nucleotide variants (SNVs) from computational analysis of next-generation re-sequencing data has proven to be relatively straightforward: sequence reads are mapped to a reference genome sequence, allowing for mismatches, and sites where the mapped reads differ in sequence from the reference are used to call variants[115, 168]. The characterization of structural variants from next-generation sequence data has

proven to be a far more challenging, but by no means intractable, problem[106]. Early methods for calling structural variants operated in manner that was agnostic with respect to the particular class of variant that was being characterized, whereas subsequent efforts have resulted in refined methods that are specifically tailored to individual structural variant classes[169, 170]. The most widely used and reliable methods for the computational detection of human TE insertion polymorphisms fall into the latter class of more specific methods[147]. We want to be clear that these novel computational methods that we are describing are aimed at the detection of TE insertion polymorphisms, which will differ from the reference genome sequence, rather than the more mature bioinformatics methods (*e.g.* RepeatMasker) that are used to characterize the identity of the more ancient, fixed TE sequences that are included as part of a reference genome sequence.

Table 9 Computational approaches for genome-wide detection of TE insertions.

Methods are sorted in order by their year of publication.

Tool Name	PMID	Year	Comments
VariationHunter	19447966	2009	Originally developed for SV detection, later refined for TE calling
HYDRA-SV	20308636	2010	General purpose SV tool; reported on mouse genome
TE-Locate	24832231	2012	Reported on 1001 Arabidopsis genomes project
Tea	22745252	2012	Specialized TE caller for cancer WGS data
ngs_te_mapper	22347367	2012	Requires TSDs; reported for Drosophila melanogaster
RetroSeq	23233656	2013	Tested on 1KGP and mouse strains
ReloaTE	23576519	2013	Requires TSDs; designed for rice genomes
Mobster	25348035	2014	Tested on 1KGP; reliable predictor for Human genome
Tangram	25228379	2014	Used in Phase II of 1KGP; no longer maintained
TEMP	24753423	2014	Reported on 1KGP and Drosophila genomes
T-lex2	25510498	2014	Reported on 1KGP and Drosophila genomes
TE-Tracker	25408240	2014	Reported on Arabidopsis genome and simulated human genome
TIGRA	24307552	2014	A breakpoint assembler and not a structural variant caller
TranspoSeq	24823667	2014	Specialized TE caller for cancer WGS data
TraFiC	25082706	2014	Specialized TE caller for cancer WGS data
MELT	26432246	2015	Used in Phase III of 1KGP; reported to work on Human, Chimp & dog.
ITIS	25887332	2015	Reported on *Medicago truncatula*; not optimized for Human genome
Jitterbug	26459856	2015	Reported on 1KGP and Arabidopsis genome
MetaSV	25861968	2015	General purpose SV tool; reported on simulated genome
DD_DETECTION	26508759	2016	Database free dispersed duplication detection approach
GRIPper	-	-	Detects non-reference gene copy insertion

There exist numerous computational tools that allow for the detection of TE insertions from next-generation whole genome sequence data (Table 9). While these programs may differ substantially in their details, they all tend to rely on the same two fundamental principles: discordant read pair mapping and split (or clipped) reads[58] (Figure 21A). Discordant read pair mapping occurs when one member of a read pair maps uniquely to the reference genome sequence and the second member of the pair maps to a repetitive TE sequence that is not found in the adjacent genomic region in the reference sequence. In some cases, the second member of the pair may map partially to unique reference genome sequence and partially to the TE sequence. The presence of multiple read pairs that show this pattern, from within the same genomic interval, is taken as evidence of a TE insertion, with the specific identity of the inserted element determined by the mapping of the second member of the read pair. Typically, a TE reference library is provided to facilitate these mappings and the corresponding characterizations of insertion identities. The discordant read pair mapping technique is ideal for short read, pair end sequencing technology, such as the Illumina SBS method. The somewhat less commonly used, at least at this time, split read technology for computational detection of polymorphic TE insertions relies on longer sequence reads that map partially to unique reference genome sequence and partially to a repetitive TE sequence. This can include reads with one end in unique genome sequence and the other end in a TE sequence or reads that span an entire TE insertion (*i.e.* have a TE sequence in the middle of the read). As longer sequence read technologies – such as the Pacific Biosciences single molecule real time sequencing method (PacBio SMRT) – become more widely used for human genome sequencing, the split read approach should become increasingly useful. Alternatively, long reads may eventually come to be used for

ab initio assembly of complex eukaryotic genomes, such as the human genome, thereby obviating the need for computational TE insertion detection methods altogether.

Two of the earliest computational methods developed specifically for the detection of TE insertions from next-generation sequence data are VariationHunter[171] and the program Spanner[59], which was used for calling TE insertions in the first phase of the 1KGP. Subsequent phases of the 1KGP included additional refinement of next-generation sequence based TE insertion calling methods resulting in the Tangram[112] and MELT[42] programs, for the second and third phases of the project, respectively. RetroSeq[62] and Mobster[111] are two of the other most widely used programs for sequence based TE insertion detection. RetroSeq was implemented primarily for the detection of endogenous retrovirus insertions in the mouse genome, whereas Mobster was tested mainly on human L1 and Alu elements.

Until very recently, all of these individual programs had only been benchmarked and validated individually by the same groups that developed each one. In other words, there was no independent and controlled comparison of the accuracy, runtime performance and usability of these tools. We recently performed just such a benchmarking and validation comparison of 21 different programs for sequence based TE insertion detection in an effort to provide researchers with an unbiased assessment of their utility[147]. Our benchmarking study was focused solely on human TE detection, owing both to the importance of human TE detection for population and clinical genetic studies as well as the availability of an experimentally validated set of TE insertions for an entire human genome.

The first phase of our benchmarking study entailed an effort to select tools that would be of the most potential use to the human TE research community. This included eliminating all programs from consideration that: 1) performed general structural variant detection (these typically have worse performance of TE insertion detection), 2) were specifically designed for cancer and required matched normal-cancer genome pairs, or 3) perform breakpoint assembly for TE insertion identification and are not able to detect insertion site locations without prior information. In this phase, we also eliminated all programs that were no longer supported and/or could not be used due to non-user generated errors such as previously reported bugs. This process resulted in reducing the original set of 21 programs down to 7 programs, which we then evaluated using simulated and actual human genome sequences.

The final 7 programs that we evaluated were: MELT[42], Mobster[111], RetroSeq[62], Tangram[112], TEMP[110], ITIS[109] and T-lex2[61] (Table 9). For each of these programs, we provided a detailed set of notes in support of their installation and use, including the exact commands and parameters that are required for their optimal performance. We compared all of the programs with respect to a set of qualitative and quantitative benchmarks. The qualitative benchmarks were ease of installation, ease of use, level of detail in the user manual and source code availability (*i.e.* open or closed source). The quantitative benchmarks were precision and recall accuracy measures along with the runtime parameters: CPUtime, walltime, RAM and the number of CPUs used. The simulated data that we used consisted of artificial genome sequences with randomly generated TE insertions and sequence read pairs simulated based on the Illumina sequencing profile. The empirical data was taken from a single individual from the 1KGP

whose genome was extensively characterized, including with PacBio long read sequence technology, resulting in an experimentally validated set of 893 TE insertions genome-wide.

When all of these factors were taken into consideration, the program MELT showed the best overall performance followed by the programs Mobster and RetroSeq. The superior performance of MELT on these particular data should be taken with some caution given the fact that it was developed and refined on the exact same human dataset. Indeed, the programs that were designed to perform more broadly, such as TEMP, or for different species, such as ITIS and T-lex2, did not perform as well, consistent with the possibility that they were at an inherent disadvantage when benchmarked on human genome sequence data from the 1KGP. Nevertheless, our benchmarking analysis clearly supports the use of MELT, and to a lesser extent Mobster and RetroSeq, for the computational detection of human TE insertions from next-generation sequence data.

There remain a number of caveats and open issues that should be considered when using these kinds of programs to predict TE insertions from whole genome sequence data. The first thing to consider is that no single method can produce optimal overall results. The best strategy is to use two or more of the top 3 methods – MELT, Mobster and RetroSeq – and then to combine the methods by looking for consensus TE calls that are supported by multiple methods. This approach has the potential effect of increasing precision at only a minor cost to recall, *i.e.* it is simultaneously conservative but can also increase the number of total TE calls by using multiple methods. Of course, a combined approach of this kind can be quite user intensive and could exceed the ability of some labs to readily implement. Perhaps the most pressing open issue regarding computational methods for TE insertion detection relates to the level of resolution at which the insertion sites can be located in the

genome sequence. In our experience, TE insertions can only be accurately localized within approximately ±100 bp. This lack of resolution makes it particularly difficult to combine results from multiple methods, as suggested above, since the same predictions will most often not be located at exactly the same genomic location. This limitation can be overcome by considering TE insertions detected within ±100 bp windows to represent the same calls. Nevertheless, further algorithm development aimed at more precise TE insertion location should prove to be an important future development in the field.

Table 10 High-throughput experimental approaches for TE insertion detection.

Next-generation sequence based methods are presented separately from methods that used tiling arrays or Sanger sequencing. Methods are sorted in descending order by their year of publication.

Next-generation sequence based			**Tiling arrays/Sanger based**		
Method	PMID	**Year**	**Method**	**PMID**	**Year**
L1-Seq	20488934	2010	TIP-Chip	20602999	2010
Transposon-Seq	20603005	2010	Fosmid-based	20602998	2010
ME-Scan	20591181	2010	AIP	22495107	2012
RC-Seq	22037309	2011			

5.4.2.2 High-throughput experimental approaches

In addition to driving the bioinformatics based efforts at TE insertion detection, next-generation sequencing techniques have also enabled a number of high-throughput experimental approaches for the detection of novel TE insertions (Table 10). Much like the computational approaches for TE detection, these high-throughput experimental techniques also share a core set of design principles[172] (Figure 21B). The first phase of these experiments consists of fragmentation of genomic DNA followed by enrichment for

sequence elements that uniquely correspond to active human TE subfamilies, mainly Alu and L1. Different methods are distinguished by the approaches that they take to genomic fragmentation as well as whether they use PCR or hybridization for the enrichment step. Enrichment of sequence fragments from active TE subfamilies is followed by next-generation sequencing, for the most recently developed methods, or hybridization to tiling arrays for some of the older methods.

The first attempt at the systematic and unbiased characterization of novel human TE insertions was based on tiling array technology and was relatively low throughput[173]. A number of next-generation sequence based techniques for TE insertion, which allowed for a substantial increase in the numbers of TE insertions that could be detected, were independently developed right around the same time in 2010 and 2011. Three such methods were published in 2010: ME-Scan[174], L1-Seq[175] and Transposon-Seq[176]. ME-Scan was used to characterize polymorphic Alu insertions, L1-Seq was applied to L1 insertions, and Transposon-Seq was used for TE insertion discovery with both families of elements. A fourth early sequence based method for TE insertion detection employed a lower-throughput approach that utilized fosmid sequences, characterized via Sanger sequencing, to characterize L1 insertions[177]. The RC-Seq method was developed in 2011 and is the only method of its kind to be applied to all three families of active human TEs: Alu, L1 and SVA. RC-Seq combines tiling array based hybridization with next-generation sequencing for TE insertion discovery.

5.5 Evolutionary genetics of active human TEs

The high-throughput approaches to TE insertion detection described in the previous section, particularly the computational genome sequence based methods, have the potential to yield genome-wide catalogs of human TE insertion polymorphisms across numerous individuals from multiple populations. The realization of this possibility is exemplified by the 1KGP, phase 3 of which includes the public release of 16,192 TE insertion genotype calls for 2,504 individuals from 26 global populations. Datasets of this kind have the potential to yield unprecedented insight into the nature of the evolutionary forces that act on TE polymorphisms.

5.5.1 Human genetic variation from TE activity

The first step in any genome-scale evolutionary analysis of human TE insertional polymorphisms involves a basic description of the nature of the genetic variation that is generated by TE activity. This includes descriptive statistics regarding the levels of TE insertion variation within and between populations along with a sense of how polymorphic TEs are distributed across the genome, particularly with respect to the location of functionally relevant genomic features such as genes and regulatory elements.

5.5.1.1 Levels and patterns of TE genetic variation

TE insertion detection programs yield presence absence genotype calls for individual loci – homozygous absent (0), heterozygous (1) and homozygous present (2) – across the entire genome, when applied to whole genome next-generation sequence data. For large scale human genome sequence initiatives, such as the 1KGP, this yields the kind of data that can be used to calculate polyTE insertion allele frequencies within and between populations. PolyTE allele frequencies (p_{TE}) can be calculated from site-specific genotype data as the total number of TE insertions observed at any given genomic site (TE_i) normalized by the total number of chromosomes in the population under consideration ($2n$): $TE_i/2n$. This can be done for individual populations or for groups of related populations, such as the 5 major continental population groups characterized as part of the 1KGP.

Population level polyTE allele frequencies can in turn be used in turn to calculate a variety of population genetic parameters that measure how genetic variation is apportioned among populations, such as heterozygosity (H) and related fixation index (F_{ST}) statistics.

$$H = 1 - (p_{TE}^2 + (1 - p_{TE})^2) \quad (1)$$

$$F_{ST} = \frac{(H_T - H_S)}{H_T} \quad (2)$$

where H_S is the sample (within population) polyTE heterozygosity and H_T is the total (between population) polyTE heterozygosity. These kinds of statistics are ideal for measuring the effects of natural selection on TE insertion polymorphisms as described later in this review.

5.5.1.2 Genomic landscape of TE insertions

Genome-wide catalogs of polyTE genotypes can also be used to systematically evaluate the landscape TE insertions and to compare their locations to the locations of functionally important genomic features such as genes, regulatory elements and epigenetic chromatin marks. The overall human TE genomic landscape is already very well defined, dating to the initial analysis of the draft human genome sequence[19] and even earlier[178-181], but the extent to which polyTE distributions resemble those of the more ancient, fixed TEs that predominate in the human genome remains an open question. When all TE-derived sequences are considered, there a number of anomalous genomic regions that are particularly enriched or depleted for human TE sequences, and these are thought to be related to gene density and tight regulatory requirements. Across the entire genome, LINE elements (L1) tend to be enriched in AT-rich DNA and are primarily found in intergenic regions, whereas SINE elements (Alu) are enriched in GC-rich DNA regions in and around gene sequences. These TE distribution patterns correspond very well to previously defined isochores[182], which are large regions of DNA with uniform GC-content patterns [183].

One particularly interesting finding from the initial analysis of the human genome sequence was that the distribution patterns of Alus change drastically for different age classes. Older subfamilies of Alus, *i.e.* those that inserted in the genome long ago, show the most skewed genomic distributions and the highest enrichment in GC-rich DNA. As the Alu subfamilies under consideration become progressively younger, they are progressively less enriched in GC-rich DNA; in fact, the very youngest AluY subfamily shows a preference for AT-rich DNA. These results were taken to indicate that Alus are preferentially retained in GC-rich DNA, and conversely more frequently lost from AT-rich DNA, since Alus are known to

insert into the AT-rich target sequences favored by L1 encoded endonucleases. This was initially thought to be due to some positive selective force acting on Alus in GC-rich DNA [18, 19], but was later shown to be more likely related to the relative ease with which Alu deletions were tolerated in gene poor AT-rich regions, compared to gene rich GC-regions where Alu deletions via ectopic recombination between nearby insertions would be far more deleterious[17, 20-24]. This issue has received substantial attention in the ensuing years and remains controversial. Now that there is a complete catalog of very recent Alu insertions, it will be very interest to see if this same patterns holds up.

5.5.2 Polymorphic TE insertions as ancestry informative markers

Ancestry informative markers (AIMs) are genetic variants that distinguish evolutionary lineages, different species or distinct populations within the same species, and can thereby be used to reconstruct evolutionary histories[129, 184]. For a number of reasons, TE insertions have proven to be extremely useful as AIMs, both within and between species[185]. Most critically, locus-specific TE insertions nearly always represent synapomorphies, *i.e.* shared derived character states that are free from homoplasies where identical states do not result from shared ancestry[3, 17, 121]. TE insertions also have the advantage that the ancestral state can be assumed to be absence of the insertion, and TE insertions are ideal AIMs for the very practical reason that they can be rapidly and accurately typed via PCR based assays.

A number of studies from the pre-genomic era used polyTE insertions to study human evolution and ancestry[47, 49, 50, 52-56]. Most of these studies have focused on Alu

elements, owing both to their relative abundance and the ease with which their shorter sequences can be PCR amplified. Far fewer studies have used L1s as AIMs, and to our knowledge, SVAs have yet to be used as markers in human evolutionary studies. Our own lab recently published the first evolutionary analysis of human populations using the genome-wide collection of human polyTE insertions characterized as part of the 1KGP[102]. These data confirmed that human polyTE insertions are substantially geographically differentiated with many population-specific insertions. Furthermore, the patterns of polyTE insertion divergence within and between populations recapitulate known patterns of human evolution. African populations show both the highest numbers of polyTE insertions and the highest levels of polyTE sequence diversity, consistent with their ancestral status. Evolutionary relationships among human populations computed from the analysis of polyTE genotypes were entirely consistent with those that have been derived from single nucleotide polymorphisms (SNPs). In addition, when select subsets of population differentiated polyTEs are used as AIMs, they were able to accurately predict patterns of human ancestry and admixture.

It is becoming increasingly apparent that patterns of human genetic ancestry and admixture are relevant to the study of human health and disease. In particular, there are numerous health disparities between human populations, and many of these are likely to be genetically based[186, 187]. Thus, the utility of polyTEs as AIMs could prove to be of clinical relevance, in applications such as admixture mapping for instance[131, 132], in addition to their applications to population genetic studies.

5.5.3 Effects of natural selection on polymorphic TE insertions

The ability to calculate polyTE allele frequencies genome-wide, as detailed in the previous section of this review, should prove to be critical for measuring the effects of natural selection on TE insertions. One aspect of natural selection on polyTE insertions is already abundantly clear: the role that negative (purifying) selection plays in eliminating deleterious insertions from the population. The fact that TE insertions are deleterious is underscored by the numerous studies that have linked TE insertions to human disease[101, 142-145]. We describe a number of such clinically related human TE studies in subsequent sections of this review. The deleterious nature of mutations generated by TE activity is not at all surprising when you consider that TE insertions can be hundreds to thousands of base pairs long. Such large-scale mutations are clearly far more substantial mutational changes than the more commonly considered SNPs. In addition, the simple fact that TE mutations are insertions of DNA sequence, rather than duplications or other re-arrangements, also attests to their potentially disruptive nature[102].

Our own previous genome-wide study of polyTE insertions turned up several lines of evidence consistent with the action of negative selection on human TEs. First of all, human polyTE insertions tend to be found at very low allele frequencies within and between human populations. Indeed, the allele frequency spectrum of polyTE insertions is highly skewed towards the lower end, and even more so than seen for SNPs, consistent with purifying selection. In addition, polyTE insertions were found to be severely under-represented in functional genomic regions including genes and exons.

As alluded to previously, the results demonstrating the action of negative selection on human polyTE insertions are not surprising considering the disruptive nature of TE insertions and their known link to diseases. In fact, it has been suggested that TEs represent such a potent mutational threat that host genomes were forced to evolve global regulatory mechanisms to repress their activity. For example, a number of epigenetic regulatory systems may have originally evolved to defend against TE activity and were only subsequently coopted to serve as host gene regulators [188, 189]. For us, it is also particularly interesting to speculate as to a possible role for positive (adaptive) selection in sweeping polyTE insertions to (relatively) high frequencies along specific population lineages. If positive selection on polyTE insertions was to be detected, it would suggest that such sequences can somehow encode functional utility for the human genome.

The possibility that TE sequences can provide functional utility for their host genomes is well supported by numerous studies on the phenomenon of exaptation[71, 190], or molecular domestication[72], whereby formerly selfish TE sequences come to encode essential cellular functions. This has been seen most often in the context of regulatory sequences[73]. Human TE sequences have been shown to provide a wide variety of gene regulatory sequences including promoters[74-76], enhancers[77-81], transcription terminators[82] and several classes of small RNAs[83-85]. Human TE sequences can also affect host gene regulation via changes in the local chromatin environment[19, 86-90]. However, all of the human TE-derived regulatory sequences studied to date correspond to relatively ancient TE insertions that are no longer capable of transposition and are consequently fixed with respect to their genomic locations. Accordingly, it is not known

whether exaptation of TE sequences can occur on the far shorter time scale that would be needed in order for polyTE insertions to show evidence of evolving by positive selection.

At this time, there are some tentative lines of evidence that are consistent with a role for positive selection in shaping the evolution of human polyTE insertions. Closer inspection of the polyTE insertion allele frequency spectrum mentioned above revealed a shift at the higher end of the spectrum, suggesting that some TE insertions may have increased in frequency owing to the effects of positive selection. This pattern was seen for Asian and European populations but not for African populations. Thus, it is possible that this shift could reflect genetic drift, and accordingly less efficacious selection, in human populations that have historically lower effective population sizes. Additional work is needed to distinguish between these two possibilities. There is also data from a more narrowly focused study on polymorphic L1 insertions showing patterns of linkage disequilibrium and extended haplotypes that are consistent with positive selection on human polyTE insertions[97].

More detailed studies on human TE genetic variation will be needed to fully assess the role that positive selection has played in the evolution of polyTEs. The flood of whole genome sequence data coming from human genome initiatives around the world, coupled with the maturing computational techniques for characterizing polyTE insertions from those data, should provide ample opportunities for studies of this kind. In addition, the analytical framework for detecting positive selection at the genomic level is already well established[191-193] and should be readily portable to genome-wide studies of TE genetic variation.

5.6 Clinical genetics of polymorphic TE insertions

5.6.1 TE insertions in Mendelian disease

Human TE insertions are relatively large scale mutations that are considered to be both rare and deleterious, particularly if they occur in genes or other functionally important genomic elements. In other words, TE insertions often correspond to highly penetrant mutations, and accordingly they have been linked to many Mendelian diseases[44, 45]. Indeed, the ability of L1 sequences to transpose was first confirmed by a study showing that a novel L1 insertion into the *F8* (Coagulation Factor VIII) gene causes hemophilia A[28]. Subsequent studies have implicated Alu insertions in a number of Mendelian diseases, including hemophilia B [63], cystic fibrosis[64] and Apert syndrome[65]. An SVA insertion in the *BTK* (*Bruton Tyrosine Kinase*) gene causes X-linked agammaglobulinaemia[66].

Despite their known disease causing properties, TE insertion mutations are often not considered in screens for disease causing variants. For example, widely used exome based methods for disease variant discovery will necessarily overlook the contribution of TE insertions to human disease. Computational and experimental approaches to TE insertion discovery provide a number of potential advantages with respect to the discovery of TE mutations that can cause Mendelian diseases. As we have discussed previously, these kinds of approaches allow for the systematic and unbiased characterization of deleterious TE insertions genome-wide, a critical dimension of genomic approaches to the diagnosis of disease. In addition, characterization of the genomic landscape of TE insertions for large scale population based genome initiatives (Table 8) will provide an important reference

panel of TE mutations that are found in healthy individuals for the purpose of screening for rare potential disease causing variants.

5.6.2 TE activity and cancer

There are a number of lines of evidence that indicate a relationship between the activity of human TEs and the etiology of cancer, particularly for the active subfamily of L1 elements. The initial studies that uncovered a potential connection between TEs and cancer focused on expression, both transcript and protein, of L1 elements in tumor tissue samples. While it was previously thought that L1 expression was largely repressed in somatic tissue, it has been shown that numerous L1 elements are also expressed in a wide variety of tumor types including testicular cancer[67], germ cell tumors[68] and breast cancer[69, 70]. More recently, nearly half of all human cancers were found to be exclusively immunoreactive to L1 *ORF1* encoded proteins compared to matched normal tissue samples, suggesting that the ORF1 proteins could serve as cancer diagnostic biomarkers[194].

In addition to the aforementioned L1 expression analysis, numerous studies have employed next-generation sequence analysis based techniques, followed by validation with PCR and Sanger sequencing, in order to characterize the TE insertion landscape of human cancers. Tumor genome sequences from a wide variety of cancer types have been found to be enriched for L1 insertions; these include colorectal tumors[146], esophageal carcinoma[195, 196] and gastrointestinal tumors[197]. In one particularly broad survey, 53% of 244 cancer genomes were found to have L1 insertions, many of which included 3' transduced sequences that are introduced as copying errors from run-on transcripts during

the reverse transcription process[198]. As was the case for TE cancer expression research, these surveys of TE insertion in cancer genomes were suggestive and interesting but did not necessarily establish a causal relationship for TE activity in the etiology of cancer (*i.e.* tumorigenesis).

A smaller number of studies have shown even more direct evidence that specific TE insertions play a causal role in the etiology of cancer. The application of the RC-Seq technique[199] to 19 hepatocellular carcinoma genome sequences uncovered two different L1 insertions, each of which initiated tumorigenesis via a different oncogenic pathway[200]. Independent L1 insertions were found in the *MCC* (*Mutated in Colorectal Cancers*) and *ST18* (*Suppression of Tumorigenicity*) tumor suppressor genes in this study. Perhaps the strongest evidence for an L1 insertion that is an actual driver mutation for tumorigenesis was recently reported for colorectal cancer[201]. Investigators in this study found a somatic L1 insertion in one allele of the *APC* (Adenomatous Polyposis Coli) tumor suppressor gene, and they showed that this L1 insertion coupled with a point mutation in the second allele of the same gene to initiate tumorigenesis via the so-called two hit colorectal cancer pathway.

5.6.3 Polymorphic TE insertion associations with common diseases

The association of TE insertions with both Mendelian disease and cancer, discussed in the previous two sections, rests on the assumptions that TE mutations are rare, deleterious and penetrant. However, recent results from analysis of the 1KGP sequences indicate that numerous TE insertions can be found in the genomes of healthy individuals[42].

Population genetic analysis of these data shown that TE polymorphisms segregate within and between human populations and can, albeit relatively rarely, increase to high allele frequencies[102]. In other words, TE polymorphisms can in some cases come to represent common genetic variants. Common genetic variants of this kind, also referred to common mutations, have been widely used over the last decade or so in association studies that aim to characterize the genetic architecture of common human diseases or conditions. Genomic characterization of TE insertion genotypes, for hundreds of thousands of individuals among various human populations, can provide an ideal source of data for genome wide association studies (GWAS), which to date have almost exclusively been conducted using SNPs.

GWAS require hundreds or thousands of cases and controls in order to have sufficient statistical power to detect associations between common genetic variants and disease. Despite the drastic decreases in the cost of whole genome sequencing over the last several years, it is still not practical to use this approach for most GWAS. Accordingly, these studies rely on the use of array technology to characterize variant alleles for hundreds of thousands of known SNPs genome-wide. This approach yields disease associations with SNP alleles that do not necessarily represent causal mutations. In other words, an associated SNP may simply tag a genomic region that contains a nearby disease causing variant that is in linkage disequilibrium (LD) with the associated SNP.

The existence of LD structure provides an important opportunity for the association of TE insertion polymorphisms with common diseases. As more and more whole genome sequences accumulate from the various genome sequencing initiatives around the world, the genomic landscape of TE insertions should become increasingly well characterized,

assuming computational methods for TE insertion detection are accurately applied to these data. The accumulation of thousands of whole genome sequences, from diverse human populations, that include genome-wide catalogs of TE insertion genotypes provides the opportunity for imputation of TE insertion genotypes via comparison with SNP array data. In this way, TE insertion polymorphisms could be associated with disease via thousands of existing GWAS studies along with untold numbers of future GWAS. The potential of this approach to TE GWAS is supported by a recent genome-wide survey of human L1 insertions that found abundant evidence of LD between these TE polymorphisms and nearby SNPs[97].

5.6.4 TE insertion associations with quantitative traits

The same logic that applies to the association of TE insertion polymorphisms with common diseases via GWAS can be used to associate TE polymorphisms with a number of different quantitative traits. These include anthropometric phenotypes, measures of human performance and a wide variety of so-called endophenotypes, which are considered as intermediate physiological traits that underlie higher order, observable phenotypes[202]. Gene expression levels are perhaps the most widely studied class of endophenotype. Expression quantitative trait loci (eQTL) analysis correlates levels of gene expression with genetic variant genotypes in order to characterize the influence of genetic variants on gene regulation. As is the case with GWAS, the vast majority of eQTL studies compare SNP genotypes with gene expression levels. However, more recent studies have begun to analyze different classes of genetic variants using the eQTL framework. For example,

copy number genotypes for short tandem repeat sequences at >2,000 loci were recently shown to be associated with the expression of numerous human genes using an eQTL approach[203]. In addition, the Structural Variation Group[42] of the 1KGP used the eQTL approach to quantify the influence of structural variants on human gene expression using RNA-seq data characterized for 1KGP samples from European and African populations by the GUEVEDIS project[155].

Many of the large scale genome initiatives listed in Table 8 will include abundant donor meta-data along with their genome sequences. For example, the NHLBI TOPMed precision medicine initiative will collect molecular, behavioral, imaging, environmental, and patient clinical data along with a variety of omics data sources, including DNA methylation, metabolite and RNA expression profiles. These kinds of quantitative data can all be compared to the genetic variants that will be characterized by whole genome sequencing, including TE insertion polymorphisms, in order to characterize the genetic architecture of a variety of quantitative human traits.

Imputation of TE genotypes onto SNP array data, as described previously in the context of GWAS, could also provide abundant opportunities to characterize TE-eQTLs in particular. The GTEx eQTL project, for instance, has compared genome-wide SNP genotypes from hundreds of individuals to their RNA-seq gene expression data for 53 human tissue types[204]. Imputation of TE insertion genotypes onto the SNP arrays used for this study could lead the discovery of TE influences on human gene expression related to a wide variety of phenotypes.

5.7 Conclusions and prospects

Human TE research has been profoundly influenced by the ongoing revolution in genomic technology. There are a number of new computational and experimental approaches that allow for the genome-wide characterization of TE insertions across numerous samples. These kinds of techniques are continually being refined and improved, and this process often goes hand-in-hand with large scale genome sequencing initiatives, such as was the case for the 1KGP. These new approaches are making it possible to study the population and clinical genetics of human TEs at the genome-scale for the first time.

The explosion of genome sequencing initiatives, which are often explicitly motivated by evolutionary or clinical considerations, will provide abundant opportunities for the application of these novel genomic techniques for TE discovery and research. Nevertheless, the sheer abundance of the data that is being generated by such initiatives will provide substantial challenges to the research community. The temptation could exist to focus on the most easily accessible sequence variants, *i.e.* SNPs, and disregard the more difficult to characterize structural variants. We feel that this would be a mistake, as it is simply not possible to appreciate the full scope of human genetic variation without considering TE insertion polymorphisms. Hopefully the new genomic technologies for TE discovery and characterization will come to be even more widely used and applied for future genome powered studies of human genetics.

APPENDIX A.

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

A.1 Notes on the general structure of the commands

Reference sequence files used in the following commands:

a) `hg19.fa` is the hg19 genome FASTA file downloaded from UCSC genome browser [154],

b) `hs37d5.fa` is the hg19 genome FASTA file with the decoy chromosome downloaded from the 1000 genomes project FTP site [42, 99].

A.2 Commands used for generating simulated BAM files

A custom written Perl script was used to spike polyTEs (AluY, L1 and SVA) in the hg19 reference genome. Following the *in silico* polyTE insertions, the insertions were verified by BLASTing the inserted sequence against the spiked reference genome and offsetting the start position to account for the inserted polyTE. The final command used was:

```
# Create the BLAST database
makeblastdb -in hg19.polyTESpiked.fa -dbtype nucl

# Query each inserted TE sequence (insertedTE.fa) against the
BLAST database
# followed by start position correction
blastn -query insertedTE.fa -db hg19.polyTESpiked.fa -
max_target_seqs 1 -outfmt "6 qseqid sseqid sstart qlen" -
```

```
num_threads 9 -perc_identity 100 -max_hsps 1 | sed 's/chr//'
| sed 's/.*_//' | awk 'BEGIN{offset=0; lastChr = 0;
OFS="\t"}{if(lastChr != $2){offset = 0; lastChr = $2}; $3 =
$3 - offset; print $2,$3-1,$3,$1; offset += $4}' >
blastOut.txt
```

Once each polyTE insertion was successfully verified, sequence reads were simulated using the ART simulator (version 2.3.7) [118]. ART was run on the Illumina MiSeq profile with the read length of 150 bp, read coverage of 5-30X, fragment length of 500bp and a standard deviation of 10bp. The following is a command for generating the simulated reads on a 5x coverage:

```
art_illumina -sam -na -i hg19.polyTESpiked.fa -l 150 -p -f 5
-s 10 -ss MS -o simulated5x -m 500
```

The output paired-end sequencing read files (simulated5x1.fq and simulated5x2.fq) were then mapped to the reference hg19 genome using bwa mem aligner [119] followed by sorting and indexing by SAMtools [120].

```
bwa mem -t 20 -I 500,10 hg19.fa simulated5x1.fq
simulated5x2.fq > simulated5x.sam
samtools sort -o simulated5x.bam -@ 10 -m 3G simulated5x.sam
&& samtools index simulated5x.bam
```

A.2 Commands used for calling polyTE in different tools

The commands listed below are general in nature and were changed for some parameters shown inside `<angular brackets>`. For generality purposes, the input bam file is always called "`map.bam`" and output prefix is "`map`" in all of the commands listed below. Often the output and the messages produced on the error streams were redirected to a log and error file.

Paired-end sequencing reads extracted from `map.bam` in FASTQ format are r1.fq and r2.fq. For the simulated data set, the original FASTQ files were available and extraction was not required. For NA12878 data sets, FASTQ reads were extracted using PICARD tools' `SamToFastq` utility.

```
java -Xmx40G -jar ~/bin/picard.jar SamToFastq I=map.bam
FASTQ=r1.fastq SECOND_END_FASTQ=r2.fastq
INCLUDE_NON_PRIMARY_ALIGNMENTS=true
```

Some of the tools used in the study contain two modes, one for detecting polyTE insertions present in the sample but absent in the reference genome and the other for detecting polyTE insertions absent in the sample but present in the reference genome. All the tools were run to detect the former mode, *i.e.*, to detect polyTE insertions absent in the reference genome.

A.2.1 MELT (Version: 1.2.20)

Dependencies: Java

java -Xmx20G -jar MELT.jar Single -h hs37d5.fa -l ./map.bam -n hg19.genes.bed -t meltTransposonFileList.txt -w output -b hs37d5 -c <coverage>

`-Xmx20G` argument controls the maximum memory of 20GB that the program (Java running MELT.jar) can use.

Files used:

1. `hg19.genes.bed` is packaged in MELT's setup add_bed_files/1KGP_Hg19/

2. meltTransposonFileList.txt contains:

```
/full/path/to/LINE_MELT.zip
/full/path/to/ALU_MELT.zip
/full/path/to/SVA_MELT.zip
```

Each of the above listed zip files is packaged in MELT's setup in me_refs/1KGP_Hg19/

A.2.2 Mobster (Version: 0.1.7c)

Dependencies: Java, Picard tools (bundled with the setup) and MOSAIK

The current available version for Mobster is 0.1.6 and requires the BAM files to be constructed using MOSAIK aligner, which wasn't the case in the data set analyzed here.

java -Xmx300G -jar MobileInsertions-1.0-SNAPSHOT.jar -properties map.properties -in map.bam -sn map -out mobster > mobster.log 2> mobster.err

The content of the `map.properties` file was mostly the same as the default properties file packaged with the Mobster program. Only three parameters were changed from the default properties file: input file name, output file name and minimum read depth coverage to suit the optimum depth of each tested data set.

Communication with the developer: Djie Tjwan Thung (January 17 – March 10, 2016)

The current released stable version (0.1.6) does not work well for alignments generated using bwa mem. The developer generously provided us the unreleased version 0.1.7c which works well with all alignments. The developer also provided us with the optimum

parameters that were used with the NA12878 high coverage data set (as specified in the properties file):

```
DISCORDANT_CLUSTER_MAX_DISTANCE=600
READS_PER_CLUSTER=1
MINIMUM_CLIP_LENGTH=35
MAXIMUM_CLIP_LENGTH=7
MINIMUM_AVG_QUALITY=20 # Different from default
READ_LENGTH=100 # Different from default
```

A.2.3 RetroSeq (Version: 1.41)

Dependencies: Perl, SAMtools, bedtools and Exonerate

The commands listed below were obtained from the RetroSeq's "1000 Genome CEU Trio Analysis" page.

Website: https://github.com/wtsi-svi/RetroSeq/wiki/1000-Genome-CEU-Trio-Analysis

```
# Discover
retroseq.pl -discover -bam map.bam -output map.bam.candidates.tab -refTEs ref_types.tab -eref probes.tab -align > log.txt 2> err.txt
# Calling phase
retroseq.pl -call -bam map.bam -input map.candidates.tab -ref hs37d5.fa -output map -filter ref_types.tab -reads <minimum read depth> -depth <maximum read depth> >> log.txt 2>> err.txt
bedtools window -b AluY_AluS.bed -a map.PE.vcf -v -w 100 > map.Alu.vcf 2>> err.txt
bedtools window -b L1HS.bed -a map.vcf -v -w 200 > map.L1.vcf 2>> err.txt
```

Minimum and maximum read depth were changed for each sample depending on the coverage of the data set.

`AluY_AluS.bed` and `L1HS.bed` comes packed with the RetroSeq package.

A.2.4 TEMP (Version: 1.04)

Dependencies: Perl, SAMtools v0.1.19, BWA, bedtools, twoBitToFa (Kent Source) and BioPerl

The command and the parameters used were obtained from the example usage in the TEMP's manual.

```
TEMP_Insertion.sh -i map.bam -s ../scripts/ -r
HomoSapienRepbaseTEConcensus.fa -t hg19_rpmk.bed -m 3 -f 500
-c 8 -u > log.txt 2> err.txt
```

The final output is in `map.insertion.refined.bp.summary`. The resulting file can be further filtered using the following commands:

```
awk 'BEGIN{OFS="\t"}{if($7 >= 5 && $8 >= 0.1){print}}'
NA12878.insertion.refined.bp.summary | cut -f1-3 | sort -k1,1
-k2,2 -V | uniq > temp.tsv
```

The "`scripts`" folder is the address of the folder containing TEMP scripts.

The files `HomoSapienRepbaseTEConcensus.fa` and `hg19_rpmk.bed` are the RepBase [117] consensus sequence and RepeatMasker [107] annotation file that comes as part of the TEMP package.

Communication with the developer: Jiali Zhuang (November 20, 2105 – March 10, 2016)

One issue with TEMP is that it only works with SAMtools version 0.1.19 or earlier. The author also recommended using the `-u` option to avoid multiple reporting of the same TE and filtering insertions that are supported by less than 5 reads (column 7 of the output) or have an allele frequency of less than 10% (column 8 of the output). For the high coverage data, we decided to use an even more stringent cut-off of 20 minimum reads and 20% allele frequency.

A.2.5 Tangram (Version: 0.3.1)

Dependencies: g++ 4.2.0+, zlib, pthread lib

Tangram's detection pipeline has multiple steps, a few which we ran parallel. These commands were derived from the manual and the usage from each program. The complete pipeline was run on the default set of parameters. Additionally, the output chromosome wise VCF files were compressed (bgzip), indexed (tabix) and concatenated (VCFtools [134]) to produce the resulting genome wide polyTE callset.

```
tangram_index -ref hs37d5.fa -sp
moblist_19Feb2010_sequence_length60.fa -out tangramIndex
tangram_bam -i map.bam -r
moblist_19Feb2010_sequence_length60.fa -o tangram.bam
samtools sort -@ 10 -m 2G tangram.bam tangramSorted
echo tangramSorted.bam > list.txt
tangram_scan -in list.txt -dir tangramScan
seq 1 22 | xargs -I CHR -P 22 sh -c 'tangram_detect -lb
tangramScan/lib_table.dat -ht tangramScan/hist.dat -in
list.txt -ref tangramIndex -rg CHR > chrCHR.vcf'
```

```
seq 1 22 | xargs -I CHR -P 22 sh -c 'bgzip chrCHR.vcf; tabix
-p vcf chrCHR.vcf.gz'
vcf-concat chr*.vcf.gz > tangram.vcf
```

The file `moblist_19Feb2010_sequence_length60.fa` was included with the Tangram package.

Communication with the developer: Jiantao Wu (March 7, 2016)

The first author and developer of the program Jiantao Wu was contacted regarding the error message "ERROR: Cannot read the number of anchors from the library file". We were not able to get any response from the developer. The same set of commands works on the low coverage data but fails on all other data sets.

A.2.6 ITIS (Download date: 1st March 2015)

Dependencies: Perl, SAMtools v0.1.19, bwa v0.7.7

The ITIS script was run on the default set of parameters.

```
itis.pl -g hg19.fa -t ./te.fasta -l 500 -N sampleName -1 r1.fq
-2 r2.fq –e Y > log.txt 2> err.txt
```

The `te.fasta` file is the FASTA consensus sequence of set of TEs expected to be polymorphic in the genome, viz., AluY, L1 and SVA. These sequences were obtained from the RepBase database [117].

Communication with the developer: Chuan Jiang (February 29 – March 11, 2016)

We were having difficulties in obtaining any predictions on any data set – actual or simulated. The developer recommended adding the `-e Y` option to the command which masks all the homologous sequences in the genome. This enabled prediction of polyTEs from the genome sequence data.

A.2.7 T-lex2 (Version: 2.2.2)

Dependencies: Perl, RepeatMasker, MAQ, SHRIMP2, BLAT

Additional dependencies for TSD: Phrap, FastaGrep

The basic command using default parameter set was selected to run T-lex2. The input files required by T-lex2 are:

1. TE list (polyTE.id; AluY, L1 and SVA)
2. TE annotations (polyTE.coord; gene coordinates for the TEs derived from RepeatMasker)
3. Reference genome (hs37d5.fa)
4. Sequencing data directory (fqDir)

```
tlex-open-v2.2.2.pl -T polyTE.id -M polyTE.coord -G hs37d5.fa
-R fqDir
```

The fqDir contained a subdirectory named after the data set being analyzed. The subdirectory contained r1.fq and r2.fq, the paired end sequence files for the respective data set.

The program took ~4 weeks to run on the low coverage human data set but did generate appropriate log messages indicating that the program was running. After ~4 weeks, the tool predicted >300,000 human polyTE insertions and was thus deemed unreliable for these particular data sets.

APPENDIX B.

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

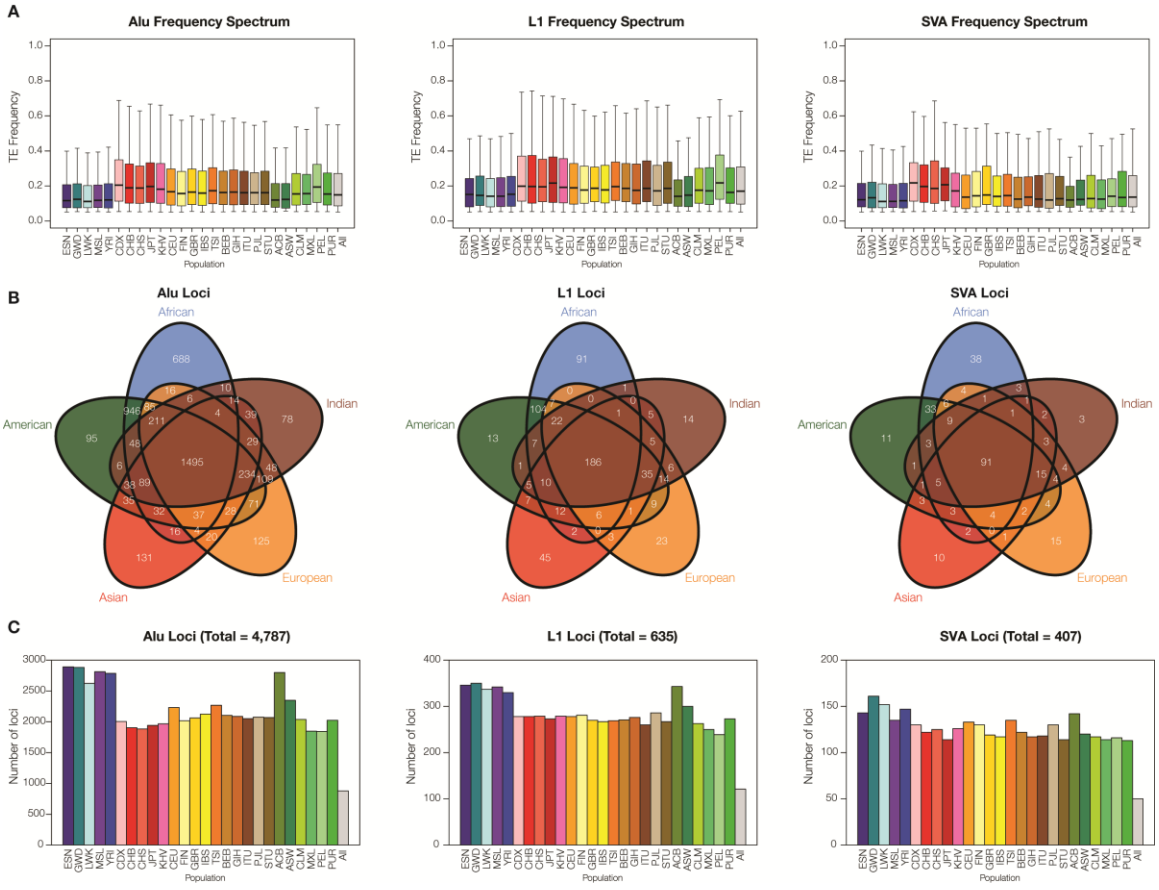

Figure 22 Distribution of polymorphic transposable element (polyTE) loci among human populations.

Data are shown individually for Alu, L1 and SVA polyTEs. Populations are organized into five continental groups (see Table 5): African (blue), Asian (red), European (gold), Indian (brown) and American (green). (A) polyTE allele frequency distribution. (B) Number of polyTE loci. (C) Numbers of shared and exclusive polyTE loci.

A

Population	Number of Samples
Brazil	45
Karitiana	24
Surui	21
Colombia	13
Colombians	13
Mexico	50
Maya	25
Pima	25
Total	108

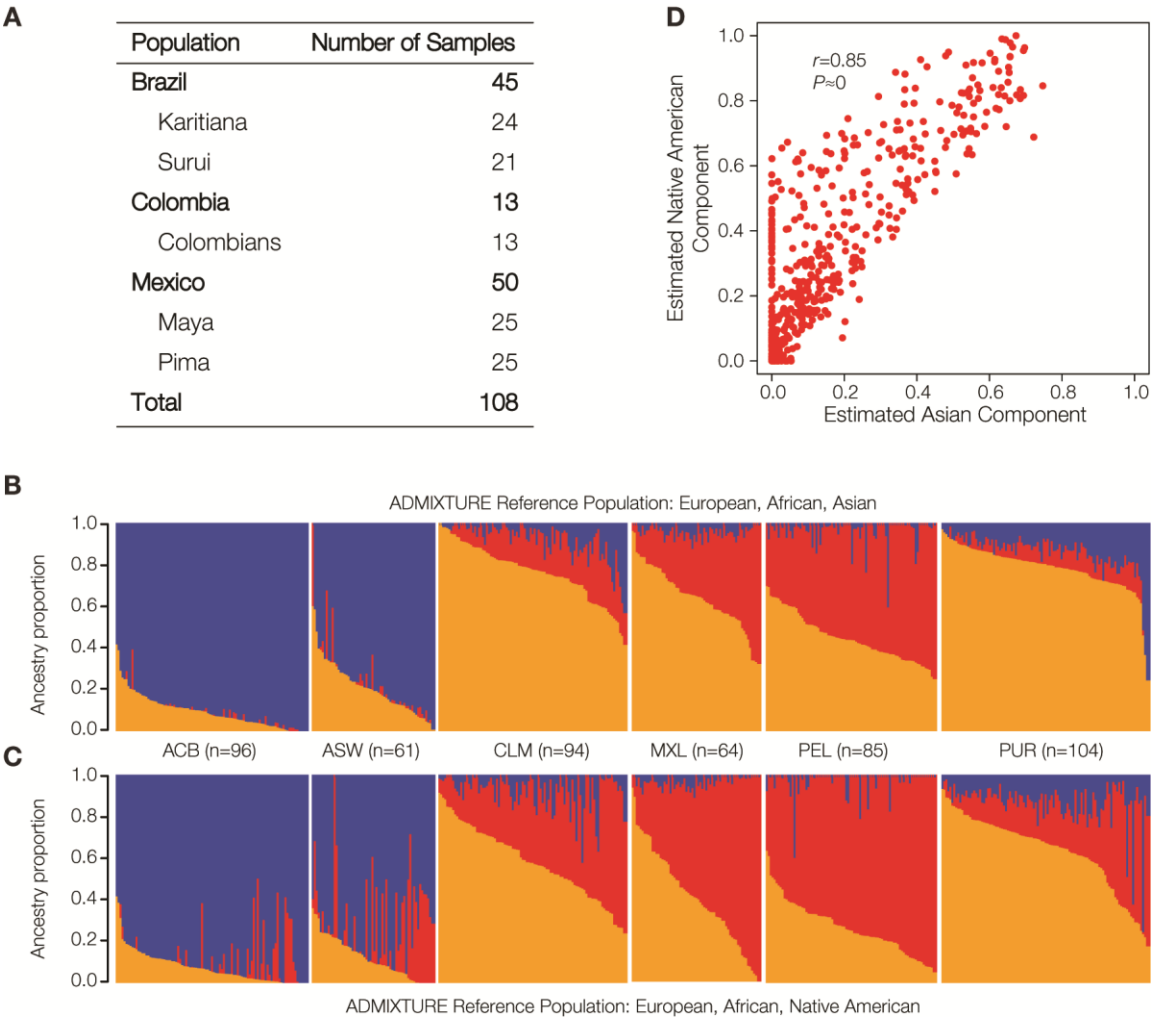

Figure 23 Continental ancestry contributions for individuals from admixed populations computed using observed Asian versus imputed Native American polyTE genotypes.

(A) Native American populations used for polyTE genotype imputation. Ancestry contribution fractions for admixed individuals computed using (B) Asian polyTE genotypes as a surrogate for Native American ancestry and (C) polyTE genotypes imputed for Native American populations. (D) Correlation between the ancestry contribution fractions computed using Asian versus Native American polyTE genotypes.

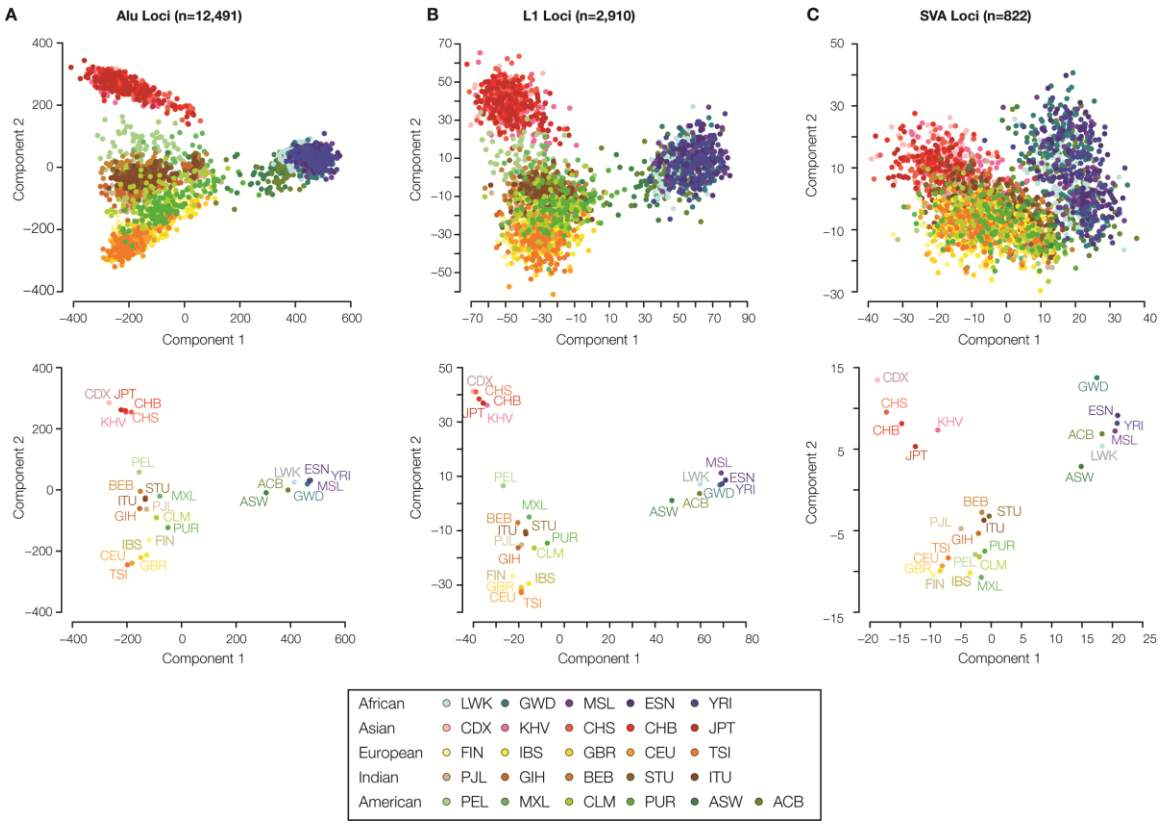

Figure 24 Clustering of human populations based on polyTE genotypes. Populations are color coded as shown in the figure legend.

Multi-dimensional scaling (MDS) plots showing polyTE genotype-based relationships among 2,504 individuals from 26 human populations are shown for (A) polyAlu, (B) polyL1 and (C) polySVA loci. The bottom panels show the same polyTE genotype MDS plots based on population average relationships.

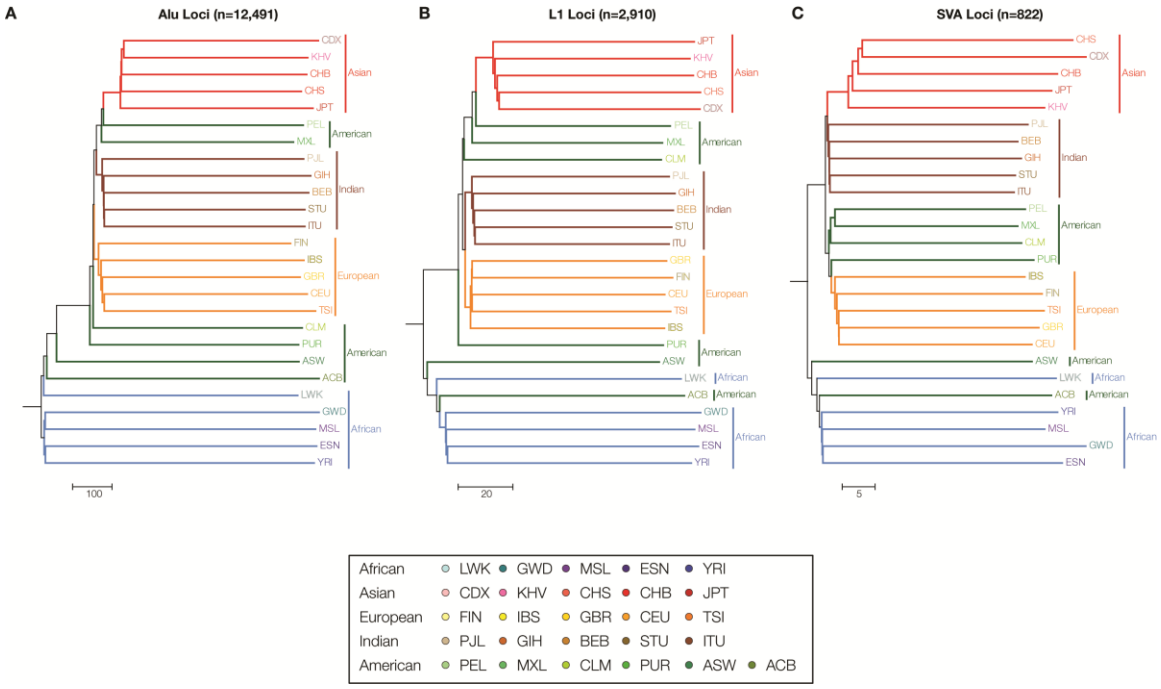

Figure 25 Phylogenetic relationships among human populations based on polyTE genotypes.

Populations are color coded as shown in the figure legend. Phylogenetic trees based on average polyTE allele sharing distances between human populations are shown for (A) polyAlu, (B) polyL1 and (C) polySVA loci.

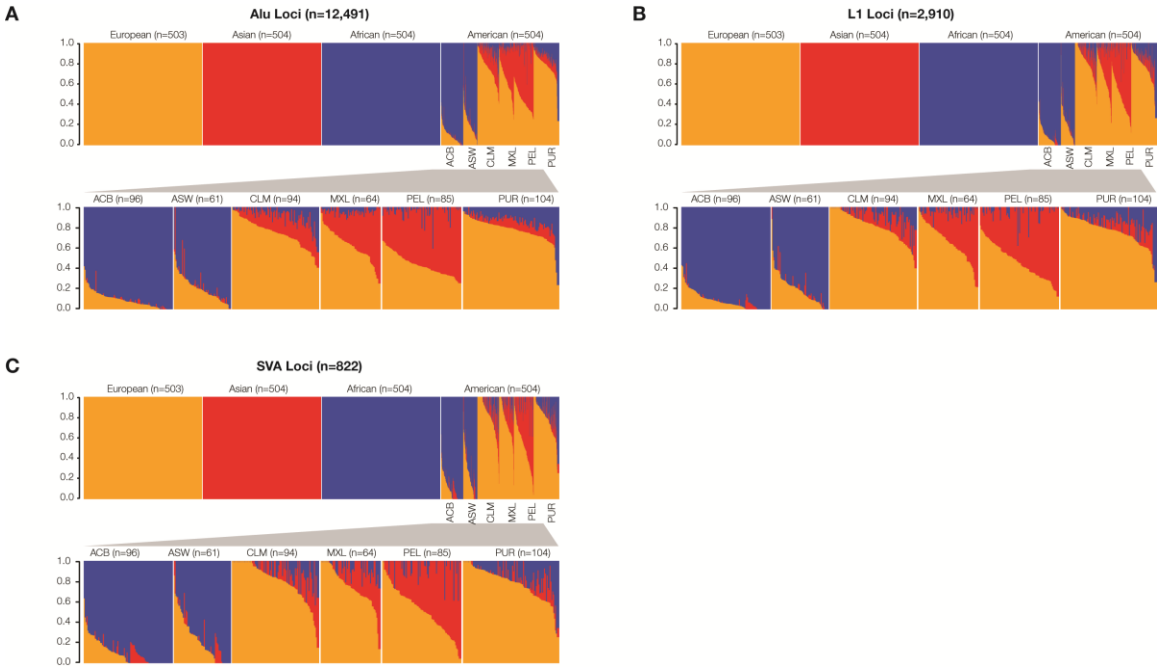

Figure 26 Continental ancestry contributions for individuals from ancestral and admixed human populations.

polyTE genotype-based continental ancestry contribution fractions for individuals from non-admixed ancestral (European, Asian and African) and admixed (American) human populations are shown for (A) polyAlu, (B) polyL1 and (C) polySVA loci.

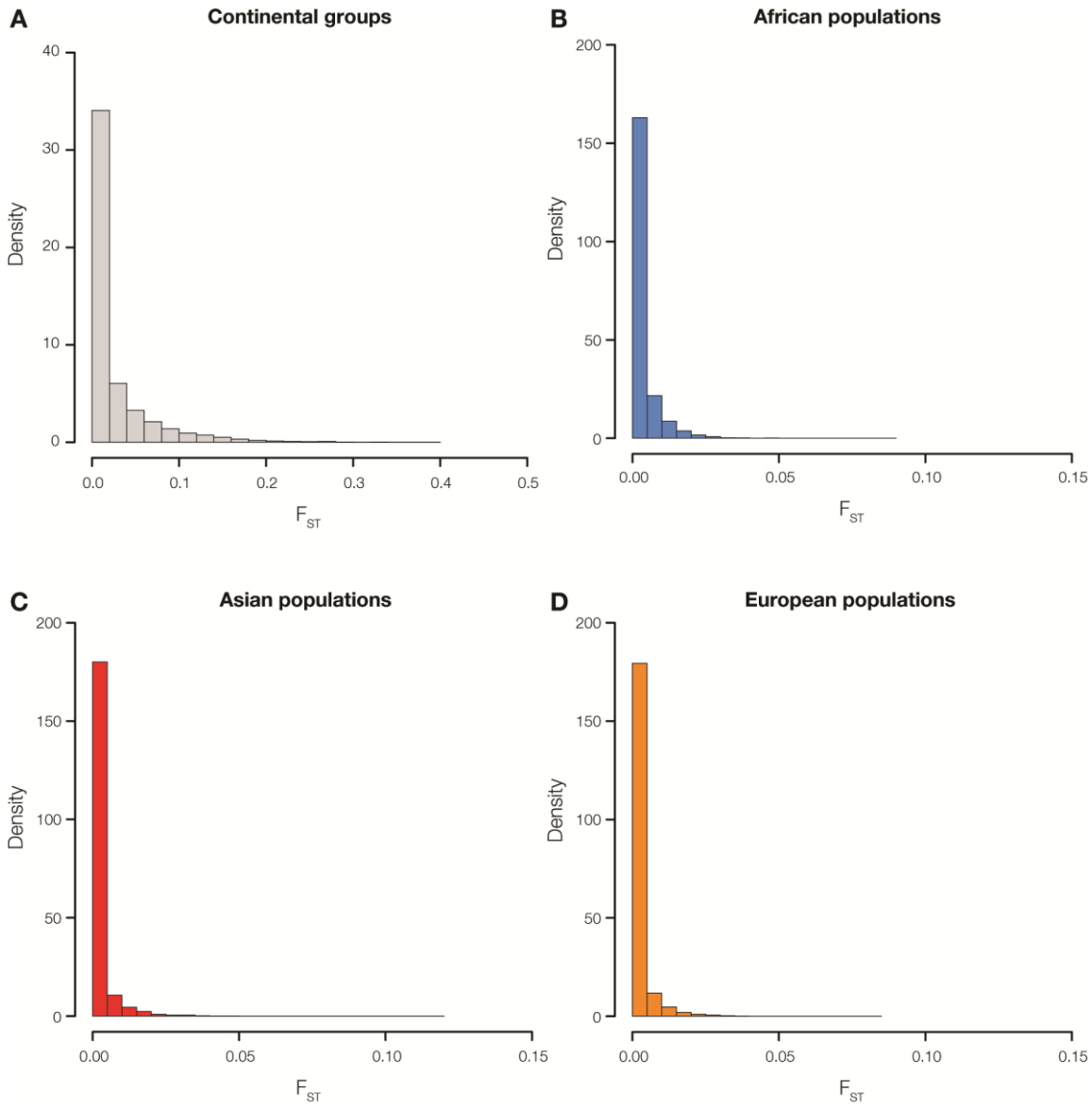

Figure 27 PolyTE genotype FST value distributions for continental group and subcontinental population comparisons.

Density distributions of F_{ST} values for polyTE genotypes are shown for (A) between continental group comparisons (African, Asian and European) and for sub-continental comparisons between (B) African, (C) Asian and (D) European populations.

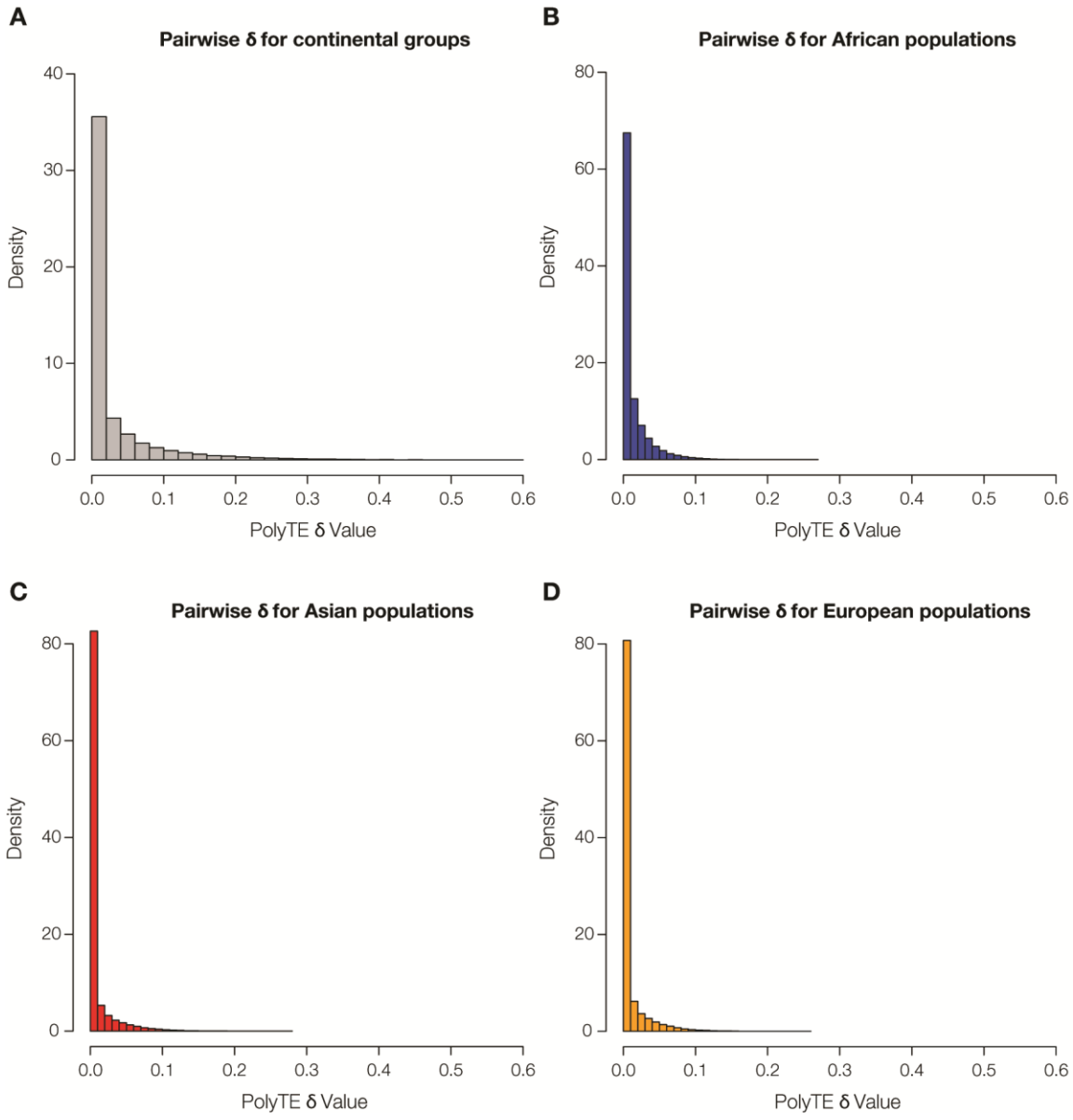

Figure 28 PolyTE genotype pairwise δ value distributions for continental groups and subcontinental population comparisons.

Density distributions of all pairwise δ values for polyTE genotypes are shown for (A) between continental group comparisons (African, Asian and European) and for sub-continental comparisons between (B) African, (C) Asian and (D) European populations.

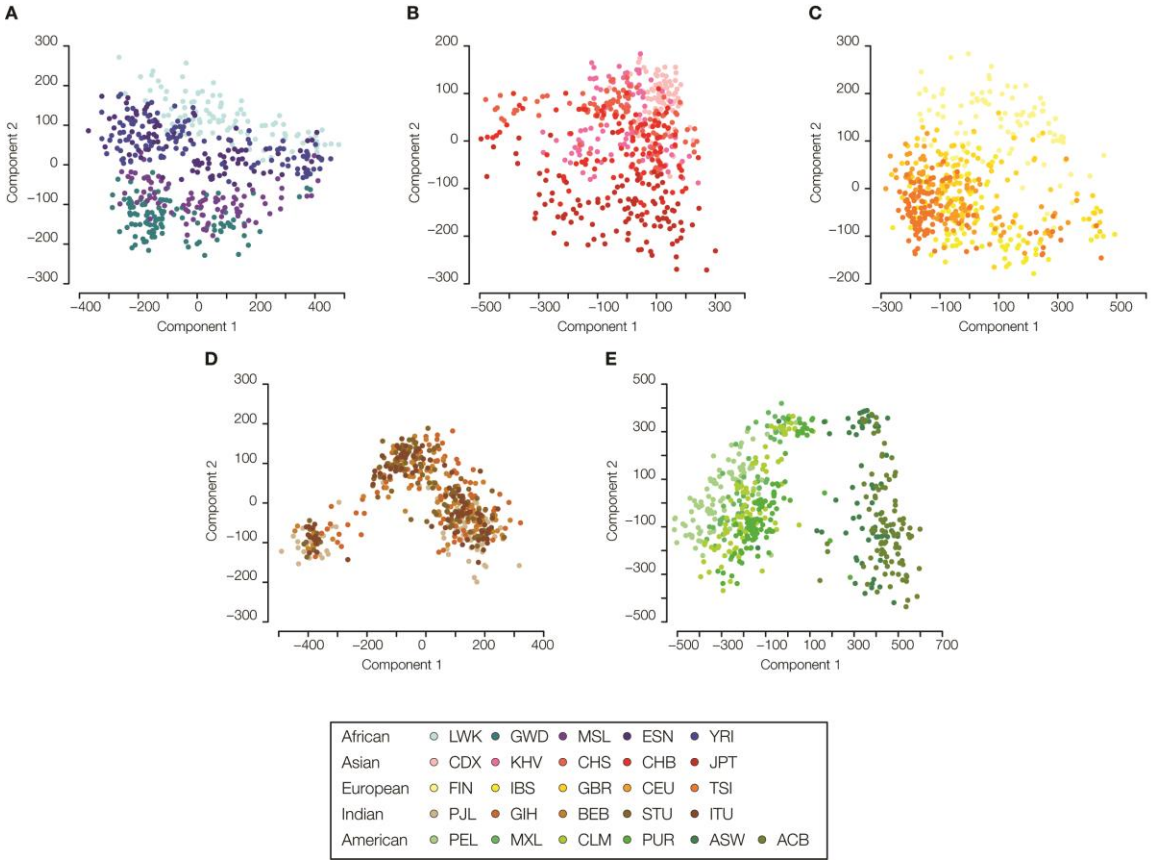

Figure 29 Sub-continental evolutionary relationships among human populations based on polyTE genotypes.

Populations are color coded as shown in the figure legend. Multi-dimensional scaling (MDS) plots showing polyTE genotype-based relationships among (A) African, (B) Asian, (C) European, (D) Indian and (E) American populations.

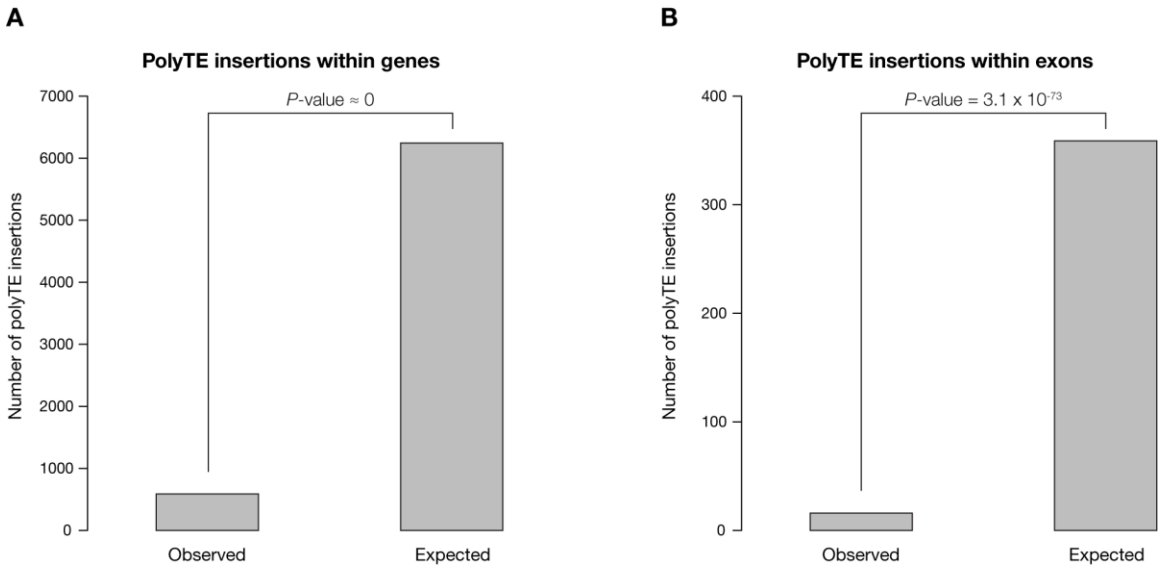

Figure 30 Numbers of polyTE insertions found within genes and exons.

The observed numbers of polyTE insertions are compared to the expected numbers for (A) gene regions and (B) exons. The expected numbers are computed based in the total number of polyTE insertions genome-wide and the fraction of the genome made up by genes (taken from Refeq transcription start to termination sites) and exons (taken from Refseq exon start to end sites). P-values are based on χ^2 tests.

Table 11 List of human polyTE loci with allele frequencies and F_{ST} values.

PolyTE Chromosomal Locations					PolyTE Allele Frequencies						PolyTE F_{ST} Values in Populations				
Chr	**Position**	**Ref**	**Type**	**ID**	**W**	**Afr**	**Asn**	**Eur**	**Ind**	**Amr**	**W**	**Afr**	**Asn**	**Eur**	**Afr vs Eur**
1	75192907	C	L1	L1_umary_LINE1_61	0.19	0.05	0.01	0.47	0.15	0.27	0.39	0.01	0.00	0.01	0.37
8	107207888	T	ALU	ALU_umary_ALU_7009	0.45	0.06	0.65	0.55	0.64	0.35	0.36	0.00	0.00	0.00	0.45
13	101577244	T	ALU	ALU_umary_ALU_10069	0.16	0.01	0.43	0.04	0.19	0.15	0.36	0.00	0.01	0.03	0.03
4	43399986	A	ALU	ALU_umary_ALU_3259	0.33	0.08	0.16	0.61	0.43	0.36	0.36	0.00	0.01	0.00	0.49
3	15294345	C	ALU	ALU_umary_ALU_2164	0.10	0.37	0.00	0.00	0.01	0.13	0.36	0.01	0.00	0.00	0.36
12	80480726	G	ALU	ALU_umary_ALU_9371	0.10	0.36	0.00	0.00	0.00	0.13	0.35	0.00	0.00	0.00	0.35
15	28175804	T	L1	L1_umary_LINE1_2628	0.24	0.00	0.54	0.20	0.25	0.20	0.35	0.00	0.01	0.00	0.19
5	37648672	A	ALU	ALU_umary_ALU_4220	0.10	0.36	0.00	0.01	0.01	0.11	0.35	0.00	0.00	0.00	0.34
4	42088056	T	ALU	ALU_umary_ALU_3251	0.44	0.07	0.65	0.51	0.55	0.40	0.34	0.01	0.01	0.00	0.38
6	126509049	G	ALU	ALU_umary_ALU_5563	0.09	0.33	0.00	0.00	0.00	0.10	0.33	0.02	0.00	0.00	0.33
2	68733422	T	ALU	ALU_umary_ALU_1278	0.09	0.34	0.00	0.01	0.00	0.09	0.33	0.00	0.00	0.00	0.32
1	75148055	T	ALU	ALU_umary_ALU_258	0.22	0.00	0.48	0.14	0.30	0.16	0.32	0.00	0.00	0.01	0.14
18	15095262	T	L1	L1_umary_LINE1_2792	0.52	0.15	0.68	0.67	0.61	0.52	0.32	0.01	0.01	0.00	0.43
16	75655176	C	ALU	ALU_umary_ALU_11116	0.19	0.05	0.44	0.04	0.24	0.19	0.32	0.00	0.00	0.00	0.00
14	57744450	A	ALU	ALU_umary_ALU_10334	0.16	0.45	0.05	0.05	0.07	0.16	0.32	0.01	0.00	0.01	0.36
2	36570238	A	L1	L1_umary_LINE1_288	0.15	0.03	0.02	0.39	0.13	0.16	0.32	0.02	0.00	0.01	0.33
17	32734421	G	ALU	ALU_umary_ALU_11284	0.36	0.03	0.52	0.50	0.45	0.31	0.31	0.01	0.00	0.00	0.45
8	50147620	G	ALU	ALU_umary_ALU_6712	0.09	0.31	0.00	0.00	0.01	0.10	0.31	0.01	0.00	0.01	0.31
7	22799818	T	ALU	ALU_umary_ALU_5885	0.27	0.22	0.55	0.04	0.30	0.25	0.31	0.01	0.00	0.00	0.13
9	13605534	G	ALU	ALU_umary_ALU_7295	0.13	0.06	0.39	0.01	0.18	0.02	0.30	0.01	0.00	0.00	0.04
3	106289201	A	ALU	ALU_umary_ALU_2591	0.27	0.10	0.58	0.16	0.31	0.22	0.30	0.01	0.00	0.00	0.01
4	74430433	T	ALU	ALU_umary_ALU_3402	0.10	0.34	0.00	0.02	0.02	0.12	0.30	0.01	0.00	0.00	0.28
5	109051004	T	ALU	ALU_umary_ALU_4562	0.22	0.04	0.49	0.14	0.30	0.13	0.30	0.00	0.01	0.00	0.06
3	21234927	G	L1	L1_umary_LINE1_558	0.10	0.32	0.01	0.00	0.06	0.09	0.29	0.00	0.00	0.00	0.31

Table 11 (continued).

10	69224896	T	L1	L1_umary_LINE1_1983	0.09	0.00	0.30	0.01	0.11	0.04	0.29	0.00	0.00	0.03	0.00
20	8816504	C	ALU	ALU_umary_ALU_12052	0.24	0.00	0.49	0.24	0.16	0.29	0.29	0.00	0.00	0.04	0.23
2	147020933	T	L1	L1_umary_LINE1_411	0.69	0.37	0.84	0.81	0.79	0.65	0.28	0.00	0.00	0.00	0.33
2	13076588	G	ALU	ALU_umary_ALU_1008	0.12	0.04	0.36	0.02	0.11	0.09	0.28	0.00	0.00	0.00	0.01
6	93576148	G	ALU	ALU_umary_ALU_5380	0.36	0.10	0.64	0.38	0.37	0.29	0.28	0.00	0.00	0.02	0.20
12	106890961	T	ALU	ALU_umary_ALU_9498	0.24	0.02	0.50	0.22	0.26	0.20	0.28	0.01	0.00	0.00	0.16
1	169442974	G	ALU	ALU_umary_ALU_581	0.11	0.03	0.35	0.02	0.09	0.04	0.28	0.03	0.02	0.01	0.00
1	158725872	T	L1	L1_umary_LINE1_142	0.08	0.28	0.00	0.00	0.00	0.09	0.28	0.00	0.00	0.00	0.28
16	82912115	A	ALU	ALU_umary_ALU_11154	0.32	0.18	0.61	0.15	0.36	0.30	0.27	0.01	0.00	0.01	0.00
5	91031026	T	ALU	ALU_umary_ALU_4463	0.10	0.32	0.02	0.01	0.05	0.12	0.27	0.02	0.00	0.00	0.29
2	192840053	C	ALU	ALU_umary_ALU_1860	0.28	0.01	0.48	0.38	0.34	0.20	0.27	0.05	0.00	0.00	0.36
1	80299106	T	ALU	ALU_umary_ALU_288	0.20	0.04	0.03	0.38	0.28	0.26	0.27	0.00	0.01	0.01	0.29
2	160159011	C	ALU	ALU_umary_ALU_1684	0.14	0.04	0.04	0.38	0.10	0.16	0.27	0.01	0.00	0.02	0.30
11	25605138	T	ALU	ALU_umary_ALU_8456	0.39	0.07	0.55	0.49	0.47	0.38	0.27	0.00	0.00	0.00	0.36
7	120103461	T	ALU	ALU_umary_ALU_6332	0.12	0.06	0.37	0.02	0.04	0.10	0.27	0.02	0.00	0.01	0.02
13	92161674	T	ALU	ALU_umary_ALU_10024	0.39	0.09	0.55	0.55	0.43	0.36	0.27	0.00	0.00	0.00	0.39
3	166544603	T	ALU	ALU_umary_ALU_2895	0.15	0.01	0.02	0.32	0.26	0.16	0.27	0.02	0.00	0.00	0.29
1	36474694	T	ALU	ALU_umary_ALU_102	0.18	0.17	0.45	0.01	0.06	0.19	0.27	0.02	0.00	0.02	0.14
14	52643767	T	ALU	ALU_umary_ALU_10309	0.23	0.06	0.47	0.12	0.27	0.23	0.27	0.00	0.00	0.01	0.02
6	103871467	C	L1	L1_umary_LINE1_1371	0.08	0.29	0.01	0.00	0.00	0.09	0.27	0.02	0.01	0.00	0.29
9	76893085	T	ALU	ALU_umary_ALU_7481	0.07	0.27	0.00	0.00	0.00	0.07	0.27	0.00	0.00	0.00	0.27
1	214911013	T	ALU	ALU_umary_ALU_780	0.12	0.35	0.00	0.07	0.03	0.13	0.27	0.00	0.00	0.00	0.22
8	122155287	T	ALU	ALU_umary_ALU_7101	0.30	0.09	0.54	0.18	0.44	0.25	0.26	0.01	0.00	0.01	0.03
8	13975433	T	ALU	ALU_umary_ALU_6584	0.23	0.09	0.49	0.11	0.29	0.16	0.26	0.00	0.01	0.00	0.00
9	105073669	T	L1	L1_umary_LINE1_1870	0.32	0.25	0.60	0.11	0.42	0.22	0.26	0.00	0.00	0.00	0.06
3	180190410	G	ALU	ALU_umary_ALU_2977	0.07	0.26	0.00	0.00	0.00	0.08	0.26	0.00	0.00	0.00	0.26
8	133725454	T	ALU	ALU_umary_ALU_7169	0.08	0.28	0.00	0.01	0.01	0.12	0.26	0.01	0.00	0.01	0.26
9	11329329	T	ALU	ALU_umary_ALU_7283	0.13	0.00	0.34	0.06	0.16	0.10	0.26	0.00	0.00	0.00	0.06

Table 11 (continued).

1	77634327	A	ALU	ALU_umary_ALU_273	0.15	0.38	0.03	0.06	0.07	0.21	0.26	0.01	0.00	0.01	0.26
4	110248315	A	L1	L1_umary_LINE1_922	0.14	0.37	0.00	0.09	0.07	0.15	0.26	0.00	0.00	0.01	0.20
12	43256994	T	L1	L1_umary_LINE1_2278	0.56	0.26	0.63	0.78	0.69	0.45	0.26	0.01	0.00	0.00	0.42
9	102849310	G	ALU	ALU_umary_ALU_7594	0.12	0.37	0.00	0.08	0.02	0.13	0.26	0.00	0.00	0.00	0.21
2	19748416	A	ALU	ALU_umary_ALU_1038	0.27	0.30	0.56	0.07	0.22	0.18	0.26	0.01	0.00	0.01	0.16
14	66934461	C	ALU	ALU_umary_ALU_10374	0.30	0.54	0.36	0.05	0.27	0.29	0.26	0.00	0.00	0.00	0.44
17	28908659	A	SVA	SVA_umary_SVA_698	0.07	0.26	0.00	0.00	0.00	0.08	0.26	0.00	0.00	0.00	0.25
5	118097178	T	ALU	ALU_umary_ALU_4615	0.07	0.26	0.00	0.00	0.00	0.08	0.26	0.00	0.00	0.01	0.25
17	43818955	C	ALU	ALU_umary_ALU_11327	0.07	0.26	0.00	0.00	0.00	0.08	0.25	0.03	0.00	0.00	0.25
7	70188702	T	L1	L1_umary_LINE1_1523	0.28	0.08	0.53	0.19	0.33	0.26	0.25	0.01	0.01	0.00	0.05
6	85947744	C	L1	L1_umary_LINE1_1348	0.06	0.00	0.25	0.00	0.01	0.03	0.25	0.00	0.01	0.00	0.00
20	15106609	T	ALU	ALU_umary_ALU_12082	0.09	0.00	0.00	0.26	0.09	0.09	0.25	0.00	0.02	0.00	0.26
19	57637028	T	ALU	ALU_umary_ALU_12001	0.09	0.01	0.31	0.03	0.07	0.06	0.25	0.00	0.00	0.02	0.02
11	67599059	G	L1	L1_umary_LINE1_2136	0.29	0.03	0.50	0.32	0.34	0.25	0.25	0.01	0.00	0.01	0.25
9	1850733	T	ALU	ALU_umary_ALU_7220	0.07	0.25	0.00	0.00	0.00	0.08	0.25	0.01	0.00	0.00	0.25
3	42898420	T	ALU	ALU_umary_ALU_2319	0.25	0.02	0.47	0.27	0.23	0.28	0.25	0.01	0.00	0.00	0.22
4	76993824	G	ALU	ALU_umary_ALU_3412	0.16	0.40	0.05	0.08	0.06	0.23	0.25	0.01	0.00	0.00	0.26
4	179444738	T	ALU	ALU_umary_ALU_3974	0.24	0.53	0.08	0.21	0.15	0.23	0.24	0.00	0.00	0.00	0.19
3	45542662	T	ALU	ALU_umary_ALU_2325	0.34	0.05	0.49	0.44	0.42	0.33	0.24	0.00	0.00	0.00	0.34
12	61247559	T	ALU	ALU_umary_ALU_9258	0.07	0.25	0.00	0.00	0.00	0.09	0.24	0.01	0.00	0.00	0.24
1	217913513	T	ALU	ALU_umary_ALU_795	0.20	0.01	0.41	0.14	0.27	0.14	0.24	0.00	0.02	0.00	0.11
1	119553366	C	L1	L1_umary_LINE1_122	0.19	0.00	0.41	0.17	0.10	0.27	0.24	0.00	0.00	0.00	0.17
13	86340152	G	L1	L1_umary_LINE1_2470	0.34	0.14	0.62	0.30	0.34	0.28	0.24	0.01	0.01	0.00	0.07
1	105138650	T	ALU	ALU_umary_ALU_421	0.15	0.07	0.41	0.07	0.18	0.04	0.24	0.00	0.00	0.00	0.00
8	69671010	T	SVA	SVA_umary_SVA_384	0.11	0.03	0.34	0.04	0.11	0.04	0.24	0.01	0.01	0.00	0.00
5	116897151	A	ALU	ALU_umary_ALU_4608	0.07	0.24	0.00	0.00	0.00	0.09	0.24	0.00	0.00	0.00	0.24
15	82035669	T	ALU	ALU_umary_ALU_10844	0.13	0.01	0.02	0.30	0.16	0.14	0.24	0.00	0.00	0.00	0.26
10	2388062	T	ALU	ALU_umary_ALU_7743	0.18	0.07	0.45	0.10	0.16	0.14	0.24	0.01	0.00	0.00	0.00

Table 11 (continued).

2	199988338	C	ALU	ALU_umary_ALU_1903	0.06	0.24	0.00	0.00	0.00	0.05	0.24	0.00	0.00	0.00	0.24
10	48419952	T	ALU	ALU_umary_ALU_7936	0.29	0.02	0.39	0.43	0.31	0.31	0.24	0.01	0.00	0.00	0.39
9	12601800	T	ALU	ALU_umary_ALU_7291	0.24	0.18	0.53	0.11	0.13	0.26	0.24	0.00	0.00	0.01	0.02
11	127667251	G	ALU	ALU_umary_ALU_8901	0.06	0.24	0.00	0.00	0.00	0.06	0.23	0.00	0.00	0.00	0.23
15	53941096	A	ALU	ALU_umary_ALU_10715	0.06	0.23	0.00	0.00	0.00	0.07	0.23	0.01	0.00	0.00	0.23
8	8920127	G	ALU	ALU_umary_ALU_6560	0.34	0.06	0.51	0.37	0.42	0.34	0.23	0.01	0.00	0.00	0.25
15	79128425	C	ALU	ALU_umary_ALU_10831	0.06	0.00	0.23	0.00	0.01	0.07	0.23	0.00	0.00	0.00	0.00
4	182418058	T	L1	L1_umary_LINE1_1035	0.06	0.24	0.00	0.00	0.00	0.05	0.23	0.01	0.00	0.00	0.23
21	21324540	T	ALU	ALU_umary_ALU_12296	0.28	0.01	0.28	0.44	0.36	0.30	0.23	0.00	0.00	0.00	0.41
3	85576571	G	L1	L1_umary_LINE1_629	0.27	0.05	0.49	0.34	0.22	0.23	0.23	0.00	0.02	0.00	0.24
8	50550379	T	ALU	ALU_umary_ALU_6716	0.10	0.32	0.03	0.03	0.01	0.12	0.23	0.01	0.01	0.01	0.25
7	42580004	T	ALU	ALU_umary_ALU_5991	0.12	0.33	0.01	0.07	0.04	0.14	0.23	0.02	0.00	0.01	0.19
8	131041235	A	L1	L1_umary_LINE1_1761	0.20	0.04	0.44	0.17	0.18	0.18	0.23	0.00	0.00	0.01	0.08
15	55129551	A	L1	L1_umary_LINE1_2652	0.06	0.23	0.00	0.00	0.00	0.06	0.23	0.01	0.00	0.00	0.23
1	49428756	T	L1	L1_umary_LINE1_25	0.17	0.42	0.14	0.04	0.07	0.20	0.23	0.01	0.00	0.00	0.34
9	22288724	T	ALU	ALU_umary_ALU_7339	0.53	0.23	0.65	0.67	0.58	0.52	0.23	0.00	0.00	0.00	0.32
1	215285291	T	ALU	ALU_umary_ALU_782	0.13	0.34	0.00	0.09	0.07	0.16	0.22	0.00	0.00	0.00	0.17
7	6314063	A	ALU	ALU_umary_ALU_5788	0.08	0.25	0.02	0.00	0.02	0.09	0.22	0.02	0.00	0.00	0.25
14	88742624	T	ALU	ALU_umary_ALU_10475	0.08	0.24	0.00	0.01	0.06	0.08	0.22	0.00	0.00	0.00	0.22
11	106044581	A	ALU	ALU_umary_ALU_8820	0.42	0.11	0.48	0.56	0.54	0.43	0.22	0.02	0.00	0.00	0.37
15	81392597	C	ALU	ALU_umary_ALU_10841	0.23	0.04	0.21	0.45	0.26	0.19	0.22	0.01	0.01	0.00	0.37
10	52540971	T	ALU	ALU_umary_ALU_7950	0.06	0.00	0.23	0.00	0.08	0.02	0.22	0.00	0.00	0.00	0.00
2	215201641	G	ALU	ALU_umary_ALU_1973	0.40	0.10	0.46	0.54	0.57	0.35	0.22	0.01	0.02	0.00	0.37
18	49301388	T	ALU	ALU_umary_ALU_11699	0.14	0.03	0.33	0.06	0.18	0.09	0.22	0.00	0.00	0.00	0.01
4	134596423	A	L1	L1_umary_LINE1_967	0.33	0.03	0.45	0.36	0.52	0.28	0.22	0.00	0.02	0.00	0.29
16	2412385	T	ALU	ALU_umary_ALU_10933	0.06	0.22	0.00	0.00	0.00	0.07	0.22	0.01	0.00	0.00	0.22
6	79047489	T	ALU	ALU_umary_ALU_5308	0.06	0.22	0.00	0.00	0.00	0.07	0.22	0.01	0.00	0.00	0.22
5	8749528	G	L1	L1_umary_LINE1_1049	0.37	0.09	0.54	0.43	0.40	0.39	0.22	0.00	0.00	0.01	0.25

Table 11 (continued).

1	78607067	T	ALU	ALU_umary_ALU_276	0.25	0.15	0.56	0.19	0.17	0.15	0.22	0.00	0.01	0.00	0.00
21	26354237	T	ALU	ALU_umary_ALU_12333	0.52	0.24	0.70	0.61	0.58	0.48	0.22	0.01	0.00	0.00	0.25
1	169524859	T	L1	L1_umary_LINE1_164	0.37	0.54	0.63	0.18	0.16	0.33	0.22	0.01	0.03	0.01	0.25
13	50912089	T	ALU	ALU_umary_ALU_9780	0.13	0.06	0.37	0.06	0.10	0.08	0.22	0.00	0.00	0.01	0.00
14	56430912	C	ALU	ALU_umary_ALU_10327	0.07	0.24	0.01	0.00	0.01	0.08	0.22	0.00	0.00	0.00	0.23
4	181503039	T	ALU	ALU_umary_ALU_3981	0.08	0.00	0.00	0.22	0.10	0.07	0.22	0.00	0.00	0.00	0.22
13	58556548	T	L1	L1_umary_LINE1_2417	0.06	0.22	0.00	0.00	0.00	0.08	0.21	0.03	0.00	0.00	0.21
10	69138296	T	ALU	ALU_umary_ALU_8063	0.07	0.00	0.22	0.00	0.12	0.03	0.21	0.00	0.00	0.00	0.00
13	82825961	T	ALU	ALU_umary_ALU_9974	0.06	0.23	0.00	0.01	0.00	0.08	0.21	0.00	0.00	0.00	0.21
15	53956115	C	L1	L1_umary_LINE1_2651	0.06	0.21	0.00	0.00	0.00	0.07	0.21	0.01	0.00	0.00	0.21
9	97314256	C	ALU	ALU_umary_ALU_7569	0.22	0.00	0.39	0.27	0.31	0.11	0.21	0.00	0.00	0.00	0.27
18	1235790	A	ALU	ALU_umary_ALU_11485	0.58	0.29	0.69	0.70	0.65	0.59	0.21	0.02	0.00	0.00	0.30
4	41598293	T	ALU	ALU_umary_ALU_3245	0.06	0.21	0.00	0.00	0.00	0.07	0.21	0.00	0.00	0.00	0.21
3	103453396	T	ALU	ALU_umary_ALU_2578	0.28	0.14	0.15	0.51	0.28	0.32	0.21	0.01	0.01	0.01	0.27
19	44377945	T	SVA	SVA_umary_SVA_759	0.06	0.21	0.00	0.00	0.00	0.08	0.21	0.01	0.00	0.00	0.21
1	79346171	G	ALU	ALU_umary_ALU_280	0.05	0.21	0.00	0.00	0.00	0.05	0.21	0.00	0.00	0.00	0.21
8	96467794	T	ALU	ALU_umary_ALU_6960	0.17	0.36	0.00	0.18	0.09	0.19	0.21	0.00	0.00	0.02	0.08
4	175211588	T	ALU	ALU_umary_ALU_3952	0.11	0.02	0.02	0.26	0.15	0.10	0.21	0.00	0.01	0.01	0.22
9	97710713	T	ALU	ALU_umary_ALU_7572	0.06	0.21	0.00	0.00	0.00	0.07	0.21	0.02	0.00	0.00	0.21
12	48312490	C	ALU	ALU_umary_ALU_9207	0.08	0.00	0.01	0.23	0.08	0.09	0.21	0.00	0.01	0.00	0.22
5	146369626	A	ALU	ALU_umary_ALU_4737	0.27	0.06	0.19	0.45	0.37	0.28	0.21	0.00	0.01	0.01	0.34
7	42603743	G	ALU	ALU_umary_ALU_5992	0.10	0.29	0.00	0.06	0.03	0.12	0.21	0.01	0.00	0.01	0.17
2	21381180	T	ALU	ALU_umary_ALU_1048	0.29	0.05	0.46	0.28	0.44	0.23	0.21	0.03	0.00	0.00	0.18
10	9474855	T	ALU	ALU_umary_ALU_7778	0.27	0.03	0.40	0.39	0.33	0.19	0.21	0.00	0.00	0.00	0.32
1	249191472	T	L1	L1_umary_LINE1_242	0.14	0.37	0.15	0.01	0.03	0.12	0.21	0.02	0.00	0.00	0.34
16	11086408	A	ALU	ALU_umary_ALU_10963	0.21	0.01	0.38	0.31	0.17	0.21	0.20	0.00	0.00	0.00	0.29
12	20908685	T	ALU	ALU_umary_ALU_9054	0.32	0.05	0.46	0.39	0.42	0.31	0.20	0.00	0.01	0.00	0.28
13	37568754	A	ALU	ALU_umary_ALU_9697	0.06	0.21	0.00	0.00	0.00	0.07	0.20	0.00	0.00	0.00	0.20

Table 11 (continued).

9	33700134	T	ALU	ALU_umary_ALU_7402	0.17	0.37	0.02	0.12	0.20	0.16	0.20	0.00	0.00	0.00	0.15
11	23932387	C	ALU	ALU_umary_ALU_8444	0.06	0.21	0.00	0.00	0.00	0.08	0.20	0.00	0.00	0.00	0.20
6	133731771	T	ALU	ALU_umary_ALU_5604	0.12	0.00	0.06	0.29	0.14	0.10	0.20	0.01	0.01	0.01	0.28
7	109283352	T	ALU	ALU_umary_ALU_6283	0.29	0.05	0.41	0.42	0.29	0.27	0.20	0.00	0.01	0.00	0.32
4	41011263	G	ALU	ALU_umary_ALU_3244	0.06	0.20	0.00	0.00	0.00	0.07	0.20	0.00	0.00	0.00	0.20
17	44153977	C	SVA	SVA_umary_SVA_706	0.07	0.00	0.00	0.21	0.05	0.10	0.20	0.00	0.00	0.04	0.20
3	98775500	T	L1	L1_umary_LINE1_651	0.07	0.22	0.00	0.01	0.04	0.07	0.20	0.00	0.00	0.01	0.19
15	39811776	T	ALU	ALU_umary_ALU_10649	0.05	0.20	0.00	0.00	0.00	0.07	0.20	0.01	0.00	0.00	0.20
13	41901883	T	ALU	ALU_umary_ALU_9723	0.06	0.20	0.00	0.00	0.00	0.08	0.20	0.00	0.00	0.00	0.20
8	141361289	C	ALU	ALU_umary_ALU_7199	0.23	0.12	0.45	0.11	0.24	0.22	0.20	0.01	0.02	0.00	0.00
11	39016741	T	ALU	ALU_umary_ALU_8530	0.05	0.20	0.00	0.00	0.00	0.06	0.20	0.00	0.00	0.00	0.20
12	93812316	T	ALU	ALU_umary_ALU_9438	0.06	0.20	0.00	0.00	0.00	0.07	0.20	0.00	0.00	0.00	0.20
8	5829282	T	L1	L1_umary_LINE1_1626	0.50	0.27	0.71	0.59	0.43	0.51	0.20	0.00	0.02	0.00	0.19
6	102846094	T	L1	L1_umary_LINE1_1370	0.10	0.00	0.28	0.06	0.05	0.08	0.20	0.00	0.00	0.03	0.06
6	73716092	A	L1	L1_umary_LINE1_1325	0.05	0.20	0.00	0.00	0.00	0.06	0.20	0.00	0.00	0.00	0.20
7	18345324	G	ALU	ALU_umary_ALU_5870	0.08	0.24	0.00	0.03	0.02	0.09	0.20	0.00	0.00	0.00	0.17
5	137130159	T	L1	L1_umary_LINE1_1215	0.15	0.00	0.31	0.10	0.23	0.12	0.20	0.00	0.00	0.01	0.10
3	67689923	T	ALU	ALU_umary_ALU_2416	0.05	0.20	0.00	0.00	0.00	0.06	0.20	0.01	0.00	0.00	0.19
12	1289984	G	ALU	ALU_umary_ALU_8943	0.05	0.20	0.00	0.00	0.00	0.06	0.20	0.00	0.00	0.01	0.19
6	72580496	C	ALU	ALU_umary_ALU_5269	0.20	0.39	0.03	0.16	0.23	0.18	0.20	0.01	0.00	0.00	0.12
22	28776052	A	ALU	ALU_umary_ALU_12467	0.06	0.00	0.00	0.20	0.04	0.07	0.20	0.00	0.00	0.00	0.20
11	22223708	T	ALU	ALU_umary_ALU_8432	0.05	0.20	0.00	0.00	0.00	0.06	0.19	0.00	0.00	0.00	0.20
20	32793460	T	ALU	ALU_umary_ALU_12140	0.16	0.39	0.11	0.05	0.10	0.15	0.19	0.02	0.01	0.00	0.28
6	110102982	T	ALU	ALU_umary_ALU_5467	0.27	0.08	0.27	0.49	0.24	0.28	0.19	0.00	0.00	0.01	0.34
5	75275331	A	ALU	ALU_umary_ALU_4383	0.08	0.02	0.00	0.22	0.06	0.10	0.19	0.01	0.00	0.00	0.18
5	168245020	G	ALU	ALU_umary_ALU_4857	0.06	0.21	0.00	0.01	0.00	0.07	0.19	0.00	0.00	0.00	0.19
8	115773126	A	L1	L1_umary_LINE1_1733	0.05	0.19	0.00	0.00	0.00	0.04	0.19	0.03	0.00	0.00	0.19
6	51739581	C	ALU	ALU_umary_ALU_5170	0.30	0.23	0.54	0.14	0.31	0.30	0.19	0.00	0.00	0.01	0.03

Table 11 (continued).

14	51432094	T	ALU	ALU_umary_ALU_10299	0.39	0.13	0.53	0.48	0.47	0.36	0.19	0.00	0.00	0.00	0.26
8	71914591	A	ALU	ALU_umary_ALU_6806	0.16	0.10	0.06	0.38	0.12	0.17	0.19	0.01	0.04	0.00	0.20
2	212925324	G	ALU	ALU_umary_ALU_1961	0.22	0.02	0.40	0.28	0.24	0.17	0.19	0.01	0.00	0.00	0.22
2	189853875	A	ALU	ALU_umary_ALU_1846	0.11	0.32	0.08	0.02	0.04	0.10	0.19	0.01	0.01	0.00	0.27
13	61462344	T	L1	L1_umary_LINE1_2422	0.50	0.31	0.76	0.52	0.45	0.46	0.19	0.02	0.01	0.01	0.09
3	120425475	T	ALU	ALU_umary_ALU_2667	0.12	0.02	0.01	0.25	0.21	0.11	0.19	0.00	0.01	0.01	0.19
1	115201401	T	ALU	ALU_umary_ALU_464	0.08	0.01	0.00	0.20	0.11	0.08	0.19	0.00	0.00	0.01	0.19
11	131361338	G	ALU	ALU_umary_ALU_8922	0.05	0.19	0.00	0.00	0.00	0.04	0.19	0.01	0.00	0.00	0.19
4	88254314	T	ALU	ALU_umary_ALU_3479	0.05	0.19	0.00	0.00	0.00	0.07	0.19	0.03	0.00	0.00	0.19
12	63569182	G	L1	L1_umary_LINE1_2301	0.13	0.32	0.07	0.03	0.11	0.12	0.19	0.00	0.00	0.00	0.25
20	59766727	T	ALU	ALU_umary_ALU_12236	0.05	0.19	0.00	0.00	0.00	0.06	0.19	0.00	0.00	0.00	0.19
2	158279991	A	ALU	ALU_umary_ALU_1674	0.13	0.01	0.33	0.11	0.15	0.06	0.19	0.00	0.01	0.01	0.07
5	55927980	A	ALU	ALU_umary_ALU_4293	0.24	0.04	0.13	0.38	0.36	0.27	0.19	0.04	0.01	0.00	0.29
21	19517313	T	L1	L1_umary_LINE1_2958	0.41	0.19	0.31	0.61	0.55	0.40	0.19	0.00	0.00	0.00	0.31
10	54682268	T	ALU	ALU_umary_ALU_7966	0.05	0.19	0.00	0.00	0.00	0.06	0.19	0.01	0.00	0.01	0.19
6	110948439	G	ALU	ALU_umary_ALU_5474	0.05	0.00	0.21	0.01	0.03	0.01	0.19	0.00	0.00	0.05	0.01
5	18570164	T	ALU	ALU_umary_ALU_4119	0.24	0.03	0.27	0.41	0.26	0.21	0.19	0.00	0.01	0.01	0.34
7	142673652	A	ALU	ALU_umary_ALU_6458	0.09	0.25	0.00	0.05	0.01	0.11	0.19	0.00	0.00	0.01	0.15
14	40882900	A	ALU	ALU_umary_ALU_10240	0.33	0.11	0.53	0.33	0.41	0.26	0.19	0.00	0.01	0.00	0.13
15	85815320	T	ALU	ALU_umary_ALU_10852	0.06	0.20	0.00	0.01	0.00	0.06	0.19	0.02	0.00	0.00	0.18
15	46045090	T	ALU	ALU_umary_ALU_10672	0.05	0.19	0.00	0.00	0.00	0.04	0.19	0.01	0.00	0.00	0.19
11	4907963	T	ALU	ALU_umary_ALU_8343	0.15	0.01	0.06	0.29	0.29	0.12	0.19	0.00	0.00	0.01	0.25
13	102389395	T	ALU	ALU_umary_ALU_10072	0.28	0.04	0.33	0.42	0.31	0.29	0.19	0.00	0.00	0.00	0.33
1	42414419	T	ALU	ALU_umary_ALU_118	0.05	0.19	0.00	0.00	0.00	0.05	0.19	0.03	0.00	0.00	0.19
13	56186576	T	ALU	ALU_umary_ALU_9803	0.49	0.26	0.67	0.61	0.46	0.48	0.19	0.01	0.00	0.00	0.22
11	63145714	T	ALU	ALU_umary_ALU_8618	0.05	0.19	0.00	0.00	0.00	0.06	0.19	0.00	0.00	0.00	0.18
12	28417298	T	ALU	ALU_umary_ALU_9107	0.09	0.00	0.03	0.23	0.12	0.07	0.19	0.00	0.02	0.00	0.23
14	39508536	T	ALU	ALU_umary_ALU_10226	0.05	0.19	0.00	0.00	0.00	0.07	0.19	0.00	0.00	0.00	0.19

Table 11 (continued).

6	46310306	G	L1	L1_umary_LINE1_1293	0.37	0.13	0.55	0.44	0.35	0.36	0.19	0.01	0.00	0.01	0.21
15	37707890	T	L1	L1_umary_LINE1_2634	0.13	0.01	0.03	0.24	0.27	0.12	0.19	0.00	0.00	0.00	0.23
5	74323809	T	ALU	ALU_umary_ALU_4379	0.05	0.19	0.00	0.00	0.00	0.06	0.19	0.00	0.00	0.00	0.18
2	226970099	T	ALU	ALU_umary_ALU_2028	0.05	0.19	0.00	0.00	0.00	0.05	0.19	0.00	0.00	0.00	0.18
8	79813676	A	L1	L1_umary_LINE1_1693	0.06	0.22	0.00	0.03	0.01	0.07	0.18	0.00	0.00	0.00	0.16
14	100422635	T	ALU	ALU_umary_ALU_10519	0.05	0.18	0.00	0.00	0.00	0.06	0.18	0.01	0.00	0.00	0.18
8	24779114	T	ALU	ALU_umary_ALU_6637	0.33	0.06	0.36	0.44	0.43	0.35	0.18	0.00	0.01	0.00	0.32
17	60434048	T	ALU	ALU_umary_ALU_11393	0.10	0.28	0.04	0.02	0.06	0.10	0.18	0.01	0.02	0.00	0.22
3	131069779	A	L1	L1_umary_LINE1_687	0.05	0.18	0.00	0.00	0.00	0.05	0.18	0.01	0.00	0.00	0.18
5	10422845	G	ALU	ALU_umary_ALU_4079	0.05	0.18	0.00	0.00	0.00	0.06	0.18	0.00	0.00	0.00	0.18
6	143951375	C	ALU	ALU_umary_ALU_5657	0.25	0.17	0.11	0.47	0.30	0.20	0.18	0.00	0.00	0.00	0.18
7	149909501	A	L1	L1_umary_LINE1_1611	0.05	0.18	0.00	0.00	0.00	0.05	0.18	0.01	0.00	0.00	0.18
5	111658690	G	ALU	ALU_umary_ALU_4581	0.14	0.04	0.33	0.08	0.16	0.11	0.18	0.00	0.00	0.00	0.02
1	152823060	T	ALU	ALU_umary_ALU_520	0.05	0.19	0.00	0.00	0.00	0.06	0.18	0.00	0.00	0.01	0.18
3	120318983	T	ALU	ALU_umary_ALU_2666	0.17	0.04	0.37	0.13	0.20	0.12	0.18	0.00	0.00	0.01	0.05
14	56606684	T	ALU	ALU_umary_ALU_10329	0.20	0.13	0.42	0.10	0.17	0.17	0.18	0.01	0.00	0.02	0.00
1	17441134	G	ALU	ALU_umary_ALU_47	0.04	0.18	0.00	0.00	0.00	0.04	0.18	0.01	0.00	0.00	0.18
6	102220590	T	ALU	ALU_umary_ALU_5428	0.18	0.12	0.37	0.05	0.24	0.11	0.18	0.00	0.00	0.00	0.03
10	58904008	A	ALU	ALU_umary_ALU_7995	0.05	0.18	0.00	0.00	0.00	0.06	0.18	0.00	0.00	0.00	0.18
5	119485630	T	ALU	ALU_umary_ALU_4625	0.05	0.18	0.00	0.00	0.00	0.06	0.18	0.01	0.00	0.00	0.18
3	181994868	T	ALU	ALU_umary_ALU_2982	0.05	0.18	0.00	0.00	0.00	0.06	0.18	0.00	0.00	0.00	0.18
11	106893671	A	ALU	ALU_umary_ALU_8823	0.06	0.20	0.00	0.01	0.03	0.06	0.18	0.01	0.00	0.00	0.17
3	157369140	G	ALU	ALU_umary_ALU_2844	0.07	0.22	0.00	0.02	0.00	0.08	0.18	0.00	0.00	0.00	0.16
4	176317571	A	ALU	ALU_umary_ALU_3957	0.05	0.18	0.00	0.00	0.00	0.08	0.18	0.02	0.00	0.00	0.18
2	79929388	A	L1	L1_umary_LINE1_343	0.05	0.18	0.00	0.00	0.00	0.06	0.18	0.01	0.00	0.00	0.18
12	25462949	G	ALU	ALU_umary_ALU_9083	0.25	0.04	0.37	0.38	0.29	0.19	0.18	0.01	0.00	0.00	0.29
13	38441116	A	ALU	ALU_umary_ALU_9704	0.05	0.18	0.00	0.00	0.00	0.05	0.18	0.01	0.00	0.00	0.18
5	155290624	C	ALU	ALU_umary_ALU_4788	0.19	0.01	0.35	0.28	0.21	0.09	0.18	0.00	0.01	0.01	0.25

Table 11 (continued).

17	3922283	T	SVA	SVA_umary_SVA_683	0.05	0.18	0.00	0.00	0.00	0.07	0.18	0.02	0.00	0.00	0.18
4	92270957	C	ALU	ALU_umary_ALU_3497	0.05	0.18	0.00	0.00	0.01	0.05	0.18	0.00	0.00	0.00	0.18
10	54681232	T	ALU	ALU_umary_ALU_7965	0.05	0.18	0.00	0.00	0.00	0.05	0.18	0.01	0.00	0.00	0.18
12	114131023	T	ALU	ALU_umary_ALU_9521	0.08	0.00	0.25	0.05	0.06	0.05	0.18	0.00	0.02	0.00	0.05
4	174044228	T	ALU	ALU_umary_ALU_3943	0.25	0.05	0.19	0.40	0.30	0.30	0.18	0.00	0.01	0.01	0.30
5	22408342	T	ALU	ALU_umary_ALU_4135	0.05	0.18	0.00	0.00	0.00	0.06	0.18	0.00	0.00	0.00	0.18
2	40613688	G	ALU	ALU_umary_ALU_1143	0.22	0.10	0.10	0.40	0.28	0.23	0.18	0.02	0.00	0.00	0.21
6	69543462	T	ALU	ALU_umary_ALU_5243	0.22	0.11	0.46	0.17	0.17	0.20	0.18	0.00	0.00	0.00	0.02
6	33030313	T	SVA	SVA_umary_SVA_282	0.34	0.43	0.57	0.16	0.26	0.27	0.18	0.02	0.04	0.02	0.17
4	1301833	G	ALU	ALU_umary_ALU_3062	0.19	0.05	0.09	0.35	0.27	0.19	0.18	0.01	0.00	0.01	0.24
12	63630483	A	ALU	ALU_umary_ALU_9274	0.21	0.41	0.15	0.08	0.25	0.16	0.18	0.00	0.00	0.00	0.26
2	20085249	G	ALU	ALU_umary_ALU_1042	0.07	0.24	0.03	0.02	0.01	0.07	0.18	0.01	0.00	0.00	0.20
11	104786608	G	ALU	ALU_umary_ALU_8813	0.41	0.17	0.56	0.52	0.45	0.35	0.18	0.00	0.00	0.01	0.24
18	5411665	T	ALU	ALU_umary_ALU_11513	0.07	0.21	0.00	0.02	0.04	0.08	0.17	0.00	0.00	0.00	0.16
4	127442204	T	ALU	ALU_umary_ALU_3681	0.05	0.18	0.00	0.00	0.00	0.05	0.17	0.01	0.00	0.00	0.17
6	56525797	A	ALU	ALU_umary_ALU_5207	0.05	0.20	0.01	0.00	0.00	0.06	0.17	0.01	0.00	0.00	0.20
2	50716664	G	ALU	ALU_umary_ALU_1193	0.05	0.18	0.00	0.00	0.00	0.06	0.17	0.01	0.00	0.00	0.17
2	158510372	G	ALU	ALU_umary_ALU_1677	0.49	0.21	0.59	0.58	0.63	0.43	0.17	0.00	0.00	0.00	0.25
18	62566687	T	ALU	ALU_umary_ALU_11768	0.05	0.01	0.20	0.00	0.04	0.01	0.17	0.00	0.05	0.00	0.00
13	92513980	A	ALU	ALU_umary_ALU_10026	0.30	0.07	0.45	0.31	0.48	0.19	0.17	0.00	0.00	0.01	0.17
4	95413280	C	ALU	ALU_umary_ALU_3517	0.13	0.33	0.08	0.05	0.06	0.13	0.17	0.00	0.00	0.01	0.23
3	114364938	A	ALU	ALU_umary_ALU_2637	0.19	0.38	0.14	0.06	0.13	0.22	0.17	0.02	0.01	0.01	0.27
18	56965826	G	ALU	ALU_umary_ALU_11734	0.05	0.00	0.19	0.01	0.04	0.03	0.17	0.00	0.00	0.03	0.01
10	2241011	G	ALU	ALU_umary_ALU_7742	0.18	0.07	0.39	0.13	0.17	0.15	0.17	0.01	0.00	0.01	0.02
9	102786689	A	ALU	ALU_umary_ALU_7593	0.08	0.25	0.00	0.05	0.02	0.09	0.17	0.00	0.00	0.00	0.14
22	50273378	T	ALU	ALU_umary_ALU_12536	0.30	0.07	0.40	0.41	0.29	0.31	0.17	0.00	0.00	0.00	0.28
5	32762578	T	ALU	ALU_umary_ALU_4195	0.18	0.01	0.35	0.24	0.17	0.14	0.17	0.01	0.01	0.00	0.21
3	148909350	A	ALU	ALU_umary_ALU_2795	0.05	0.18	0.00	0.00	0.00	0.06	0.17	0.00	0.00	0.00	0.17

Table 11 (continued).

6	91133173	A	ALU	ALU_umary_ALU_5367	0.04	0.18	0.00	0.00	0.00	0.05	0.17	0.00	0.00	0.00	0.17
5	59619679	T	ALU	ALU_umary_ALU_4311	0.05	0.17	0.00	0.00	0.00	0.06	0.17	0.00	0.00	0.00	0.17
16	26101608	C	ALU	ALU_umary_ALU_11012	0.06	0.00	0.19	0.01	0.04	0.07	0.17	0.00	0.02	0.01	0.01
18	50230474	A	ALU	ALU_umary_ALU_11704	0.05	0.17	0.00	0.00	0.00	0.06	0.17	0.02	0.00	0.00	0.17
12	85506293	A	L1	L1_umary_LINE1_2328	0.17	0.32	0.00	0.24	0.07	0.20	0.17	0.00	0.00	0.03	0.01
10	19786309	T	ALU	ALU_umary_ALU_7824	0.17	0.38	0.11	0.08	0.10	0.20	0.17	0.00	0.00	0.00	0.23
3	40241622	T	ALU	ALU_umary_ALU_2306	0.28	0.08	0.46	0.29	0.34	0.22	0.17	0.01	0.01	0.00	0.13
13	49336673	G	ALU	ALU_umary_ALU_9769	0.05	0.17	0.00	0.00	0.00	0.07	0.17	0.01	0.00	0.00	0.17
8	124680183	T	ALU	ALU_umary_ALU_7114	0.32	0.09	0.41	0.46	0.32	0.32	0.17	0.00	0.01	0.01	0.28
6	43895487	T	ALU	ALU_umary_ALU_5120	0.46	0.22	0.61	0.56	0.49	0.41	0.17	0.00	0.00	0.00	0.22
7	107829263	A	L1	L1_umary_LINE1_1566	0.55	0.34	0.76	0.59	0.55	0.52	0.17	0.00	0.00	0.00	0.12
15	35473633	T	ALU	ALU_umary_ALU_10622	0.24	0.11	0.46	0.18	0.24	0.23	0.17	0.00	0.00	0.01	0.02
4	130812825	C	ALU	ALU_umary_ALU_3700	0.04	0.17	0.00	0.00	0.00	0.05	0.17	0.00	0.00	0.00	0.17
22	23872686	G	ALU	ALU_umary_ALU_12453	0.25	0.05	0.30	0.41	0.34	0.18	0.17	0.01	0.00	0.01	0.31
10	80596636	T	ALU	ALU_umary_ALU_8100	0.04	0.17	0.00	0.00	0.00	0.04	0.17	0.00	0.00	0.00	0.17
1	191817437	C	ALU	ALU_umary_ALU_694	0.14	0.01	0.32	0.14	0.12	0.09	0.17	0.00	0.02	0.00	0.11
9	30674468	T	L1	L1_umary_LINE1_1818	0.19	0.42	0.12	0.12	0.09	0.18	0.17	0.00	0.00	0.01	0.21
17	15043040	A	ALU	ALU_umary_ALU_11230	0.05	0.17	0.00	0.00	0.00	0.06	0.17	0.00	0.00	0.00	0.16
10	122662191	C	L1	L1_umary_LINE1_2031	0.05	0.17	0.00	0.00	0.00	0.06	0.17	0.01	0.00	0.00	0.17
2	126051106	T	ALU	ALU_umary_ALU_1497	0.07	0.22	0.03	0.01	0.01	0.07	0.17	0.00	0.00	0.00	0.20
14	59160899	T	L1	L1_umary_LINE1_2570	0.05	0.17	0.00	0.00	0.00	0.06	0.17	0.00	0.00	0.00	0.17
1	100994221	G	ALU	ALU_umary_ALU_402	0.62	0.35	0.72	0.70	0.71	0.63	0.17	0.00	0.01	0.00	0.22
20	5055244	C	ALU	ALU_umary_ALU_12032	0.05	0.17	0.00	0.00	0.01	0.04	0.17	0.00	0.00	0.00	0.16
7	121122564	A	ALU	ALU_umary_ALU_6339	0.04	0.17	0.00	0.00	0.00	0.05	0.17	0.00	0.00	0.00	0.16
13	79774288	A	ALU	ALU_umary_ALU_9949	0.13	0.06	0.32	0.06	0.10	0.11	0.17	0.01	0.00	0.00	0.00
10	3569025	T	ALU	ALU_umary_ALU_7750	0.33	0.10	0.33	0.49	0.48	0.28	0.17	0.00	0.01	0.01	0.30
3	187558301	G	ALU	ALU_umary_ALU_2999	0.05	0.17	0.00	0.00	0.00	0.06	0.17	0.00	0.00	0.00	0.16
1	245175788	T	ALU	ALU_umary_ALU_928	0.04	0.17	0.00	0.00	0.00	0.05	0.17	0.00	0.00	0.00	0.16

Table 11 (continued).

4	86470382	A	ALU	ALU_umary_ALU_3474	0.37	0.31	0.17	0.56	0.41	0.40	0.17	0.01	0.00	0.00	0.12
17	14417928	C	ALU	ALU_umary_ALU_11223	0.04	0.17	0.00	0.00	0.00	0.05	0.17	0.00	0.00	0.00	0.16
12	22123796	T	ALU	ALU_umary_ALU_9064	0.54	0.27	0.62	0.64	0.66	0.52	0.16	0.00	0.00	0.00	0.24
5	56830734	T	ALU	ALU_umary_ALU_4298	0.13	0.32	0.12	0.03	0.11	0.09	0.16	0.00	0.02	0.01	0.27
13	66475612	T	ALU	ALU_umary_ALU_9872	0.04	0.16	0.00	0.00	0.00	0.04	0.16	0.00	0.00	0.00	0.16
11	41530248	T	ALU	ALU_umary_ALU_8545	0.14	0.30	0.00	0.17	0.08	0.15	0.16	0.00	0.00	0.00	0.05
8	1310538	T	ALU	ALU_umary_ALU_6532	0.10	0.25	0.01	0.05	0.08	0.08	0.16	0.01	0.00	0.01	0.14
3	122158366	G	ALU	ALU_umary_ALU_2671	0.04	0.00	0.17	0.00	0.01	0.00	0.16	0.00	0.01	0.00	0.00
19	27836578	A	ALU	ALU_umary_ALU_11912	0.05	0.00	0.00	0.17	0.03	0.06	0.16	0.00	0.00	0.00	0.16
5	75874710	C	ALU	ALU_umary_ALU_4386	0.09	0.02	0.24	0.04	0.08	0.06	0.16	0.01	0.00	0.00	0.01
3	137565396	T	ALU	ALU_umary_ALU_2733	0.09	0.25	0.00	0.07	0.01	0.10	0.16	0.00	0.00	0.02	0.11
8	83818122	C	ALU	ALU_umary_ALU_6892	0.04	0.16	0.00	0.00	0.00	0.05	0.16	0.01	0.00	0.00	0.16
1	90914512	G	L1	L1_umary_LINE1_84	0.10	0.11	0.29	0.01	0.01	0.09	0.16	0.01	0.04	0.03	0.09
11	9104553	T	SVA	SVA_umary_SVA_487	0.14	0.01	0.33	0.18	0.11	0.09	0.16	0.00	0.01	0.00	0.15
10	23038017	T	ALU	ALU_umary_ALU_7838	0.05	0.16	0.00	0.00	0.00	0.06	0.16	0.02	0.00	0.00	0.16
13	61367408	G	ALU	ALU_umary_ALU_9838	0.21	0.04	0.38	0.24	0.13	0.25	0.16	0.00	0.01	0.00	0.15
11	25108568	T	ALU	ALU_umary_ALU_8453	0.07	0.21	0.01	0.02	0.06	0.06	0.16	0.00	0.02	0.00	0.15
9	12306476	A	ALU	ALU_umary_ALU_7289	0.06	0.20	0.02	0.01	0.02	0.06	0.16	0.02	0.00	0.00	0.18
15	47507342	A	L1	L1_umary_LINE1_2640	0.29	0.46	0.09	0.32	0.24	0.32	0.16	0.00	0.01	0.00	0.04
21	33385567	T	ALU	ALU_umary_ALU_12381	0.04	0.16	0.00	0.00	0.00	0.04	0.16	0.00	0.00	0.00	0.16
2	26623683	T	ALU	ALU_umary_ALU_1074	0.18	0.12	0.01	0.30	0.27	0.18	0.16	0.00	0.01	0.00	0.09
2	212845799	T	ALU	ALU_umary_ALU_1960	0.13	0.01	0.24	0.05	0.24	0.11	0.16	0.00	0.01	0.00	0.04
8	112311307	T	ALU	ALU_umary_ALU_7049	0.05	0.16	0.00	0.00	0.00	0.06	0.16	0.00	0.00	0.00	0.16
2	210260754	A	ALU	ALU_umary_ALU_1947	0.13	0.04	0.01	0.23	0.25	0.14	0.16	0.01	0.01	0.00	0.14
7	26794918	A	L1	L1_umary_LINE1_1474	0.04	0.16	0.00	0.00	0.00	0.05	0.16	0.00	0.00	0.00	0.16
11	99502919	T	L1	L1_umary_LINE1_2169	0.04	0.00	0.16	0.00	0.03	0.00	0.16	0.00	0.00	0.00	0.00
5	76716209	G	ALU	ALU_umary_ALU_4390	0.11	0.00	0.27	0.11	0.08	0.08	0.16	0.00	0.01	0.00	0.10
6	146881725	T	ALU	ALU_umary_ALU_5676	0.04	0.16	0.00	0.00	0.00	0.05	0.16	0.00	0.00	0.00	0.16

Table 11 (continued).

7	154858220	T	ALU	ALU_umary_ALU_6505	0.30	0.13	0.19	0.47	0.38	0.32	0.16	0.00	0.00	0.02	0.24
11	76990585	T	ALU	ALU_umary_ALU_8650	0.09	0.01	0.06	0.24	0.06	0.10	0.16	0.03	0.02	0.00	0.22
3	115249807	T	ALU	ALU_umary_ALU_2644	0.04	0.16	0.00	0.00	0.00	0.05	0.16	0.01	0.00	0.00	0.16
1	208691498	T	ALU	ALU_umary_ALU_758	0.18	0.22	0.36	0.03	0.14	0.12	0.16	0.00	0.00	0.00	0.14
11	41869874	A	L1	L1_umary_LINE1_2105	0.04	0.16	0.00	0.00	0.00	0.05	0.16	0.00	0.00	0.00	0.15
6	47708271	G	ALU	ALU_umary_ALU_5145	0.24	0.10	0.45	0.22	0.11	0.30	0.16	0.00	0.00	0.00	0.05
2	195969841	A	ALU	ALU_umary_ALU_1876	0.04	0.16	0.00	0.00	0.00	0.05	0.15	0.02	0.00	0.00	0.15
17	58992972	T	ALU	ALU_umary_ALU_11384	0.04	0.15	0.00	0.00	0.00	0.05	0.15	0.01	0.00	0.00	0.15
2	33404736	G	ALU	ALU_umary_ALU_1103	0.04	0.16	0.00	0.00	0.00	0.05	0.15	0.01	0.00	0.00	0.15
11	9758001	T	ALU	ALU_umary_ALU_8374	0.15	0.28	0.27	0.01	0.04	0.16	0.15	0.00	0.01	0.00	0.27
5	163311132	T	ALU	ALU_umary_ALU_4837	0.04	0.16	0.00	0.00	0.00	0.04	0.15	0.01	0.00	0.00	0.15
4	79159515	T	ALU	ALU_umary_ALU_3428	0.45	0.18	0.53	0.52	0.62	0.42	0.15	0.00	0.00	0.00	0.22
4	140647797	T	L1	L1_umary_LINE1_976	0.03	0.00	0.15	0.00	0.00	0.00	0.15	0.00	0.01	0.00	0.00
13	33305690	T	ALU	ALU_umary_ALU_9675	0.05	0.17	0.00	0.00	0.01	0.05	0.15	0.01	0.00	0.00	0.16
2	13620731	T	ALU	ALU_umary_ALU_1011	0.04	0.15	0.00	0.00	0.00	0.05	0.15	0.01	0.00	0.00	0.15
3	112734445	G	ALU	ALU_umary_ALU_2626	0.03	0.00	0.16	0.00	0.01	0.00	0.15	0.00	0.01	0.01	0.00
3	169270604	T	ALU	ALU_umary_ALU_2911	0.06	0.21	0.01	0.02	0.00	0.07	0.15	0.01	0.02	0.01	0.15
14	49672288	T	ALU	ALU_umary_ALU_10291	0.05	0.18	0.01	0.00	0.02	0.05	0.15	0.00	0.00	0.00	0.17
10	69980006	C	SVA	SVA_umary_SVA_449	0.04	0.15	0.00	0.00	0.00	0.04	0.15	0.01	0.00	0.00	0.15
4	31444954	T	ALU	ALU_umary_ALU_3191	0.09	0.23	0.00	0.06	0.05	0.10	0.15	0.02	0.00	0.00	0.11
10	1292107	G	ALU	ALU_umary_ALU_7740	0.20	0.00	0.29	0.23	0.36	0.13	0.15	0.01	0.01	0.00	0.22
13	101751424	T	ALU	ALU_umary_ALU_10070	0.05	0.17	0.00	0.00	0.02	0.05	0.15	0.00	0.00	0.01	0.15
7	24390238	T	ALU	ALU_umary_ALU_5897	0.04	0.16	0.00	0.00	0.00	0.04	0.15	0.01	0.00	0.00	0.15
3	144610083	T	L1	L1_umary_LINE1_704	0.06	0.20	0.00	0.03	0.01	0.07	0.15	0.02	0.00	0.00	0.13
8	137969623	C	ALU	ALU_umary_ALU_7182	0.04	0.15	0.00	0.00	0.01	0.04	0.15	0.00	0.00	0.00	0.15
2	154100588	A	L1	L1_umary_LINE1_423	0.05	0.16	0.00	0.00	0.00	0.07	0.15	0.00	0.00	0.00	0.15
18	2337133	T	ALU	ALU_umary_ALU_11498	0.04	0.15	0.00	0.00	0.00	0.04	0.15	0.00	0.00	0.00	0.15
10	29146319	T	ALU	ALU_umary_ALU_7875	0.14	0.32	0.14	0.02	0.09	0.12	0.15	0.00	0.03	0.00	0.26

Table 11 (continued).

2	78251635	T	ALU	ALU_umary_ALU_1323	0.08	0.23	0.01	0.04	0.02	0.09	0.15	0.01	0.05	0.00	0.14
1	239426267	G	ALU	ALU_umary_ALU_897	0.17	0.36	0.06	0.14	0.11	0.21	0.15	0.01	0.00	0.01	0.12
6	102674404	T	ALU	ALU_umary_ALU_5432	0.05	0.16	0.01	0.00	0.01	0.04	0.15	0.00	0.00	0.00	0.16
6	93032758	T	ALU	ALU_umary_ALU_5378	0.05	0.00	0.16	0.01	0.01	0.08	0.15	0.00	0.01	0.02	0.00
10	59183386	C	ALU	ALU_umary_ALU_7997	0.04	0.15	0.00	0.00	0.00	0.05	0.15	0.00	0.00	0.00	0.15
20	22501649	T	L1	L1_umary_LINE1_2917	0.04	0.15	0.00	0.00	0.00	0.04	0.15	0.01	0.00	0.00	0.15
11	81283884	T	ALU	ALU_umary_ALU_8672	0.17	0.12	0.39	0.10	0.09	0.14	0.15	0.00	0.00	0.01	0.00
5	64023155	C	ALU	ALU_umary_ALU_4335	0.04	0.15	0.00	0.00	0.00	0.04	0.15	0.00	0.00	0.00	0.15
8	95794736	T	ALU	ALU_umary_ALU_6957	0.05	0.17	0.00	0.01	0.01	0.06	0.15	0.00	0.00	0.02	0.14
4	178789752	T	ALU	ALU_umary_ALU_3968	0.03	0.00	0.15	0.00	0.01	0.00	0.15	0.00	0.00	0.00	0.00
22	42135565	A	ALU	ALU_umary_ALU_12510	0.04	0.15	0.00	0.00	0.00	0.04	0.15	0.00	0.00	0.00	0.14
3	84832133	A	ALU	ALU_umary_ALU_2497	0.12	0.10	0.28	0.02	0.06	0.13	0.15	0.01	0.00	0.00	0.04
15	23214982	C	ALU	ALU_umary_ALU_10554	0.04	0.15	0.00	0.00	0.00	0.04	0.15	0.03	0.00	0.00	0.14
6	158250192	T	ALU	ALU_umary_ALU_5723	0.04	0.15	0.00	0.00	0.00	0.05	0.15	0.01	0.00	0.00	0.14
1	116983571	T	ALU	ALU_umary_ALU_469	0.06	0.19	0.01	0.02	0.04	0.07	0.14	0.00	0.01	0.00	0.15
4	83340696	C	ALU	ALU_umary_ALU_3457	0.13	0.01	0.30	0.17	0.09	0.07	0.14	0.00	0.00	0.00	0.13
14	27526157	C	ALU	ALU_umary_ALU_10162	0.06	0.19	0.03	0.00	0.00	0.07	0.14	0.00	0.02	0.00	0.19
8	107488264	T	ALU	ALU_umary_ALU_7013	0.04	0.15	0.00	0.00	0.00	0.04	0.14	0.01	0.00	0.00	0.14
9	80365050	C	ALU	ALU_umary_ALU_7505	0.04	0.16	0.00	0.01	0.00	0.05	0.14	0.01	0.00	0.01	0.14
19	5393852	T	ALU	ALU_umary_ALU_11866	0.04	0.14	0.00	0.00	0.00	0.04	0.14	0.01	0.00	0.00	0.14
3	109728160	A	ALU	ALU_umary_ALU_2607	0.03	0.14	0.00	0.00	0.00	0.03	0.14	0.00	0.00	0.00	0.14
4	88305920	C	SVA	SVA_umary_SVA_219	0.04	0.14	0.00	0.00	0.00	0.03	0.14	0.01	0.00	0.00	0.14
3	85868023	C	L1	L1_umary_LINE1_631	0.04	0.15	0.00	0.00	0.00	0.07	0.14	0.00	0.00	0.00	0.14
1	247865134	A	ALU	ALU_umary_ALU_949	0.04	0.15	0.00	0.00	0.00	0.05	0.14	0.00	0.01	0.00	0.15
1	99894544	A	ALU	ALU_umary_ALU_393	0.04	0.14	0.00	0.00	0.00	0.05	0.14	0.01	0.00	0.00	0.14
4	179438566	T	ALU	ALU_umary_ALU_3973	0.04	0.00	0.14	0.00	0.01	0.03	0.14	0.00	0.00	0.00	0.00
6	129045852	C	ALU	ALU_umary_ALU_5575	0.19	0.25	0.38	0.06	0.05	0.20	0.14	0.01	0.00	0.00	0.14
9	80089958	T	ALU	ALU_umary_ALU_7503	0.05	0.16	0.00	0.01	0.02	0.06	0.14	0.00	0.00	0.00	0.14

Table 11 (continued).

8	53173030	A	ALU	ALU_umary_ALU_6730	0.04	0.14	0.00	0.00	0.00	0.04	0.14	0.00	0.00	0.00	0.14
2	117187141	T	ALU	ALU_umary_ALU_1459	0.04	0.14	0.00	0.00	0.00	0.05	0.14	0.00	0.00	0.00	0.14
6	58481778	T	L1	L1_umary_LINE1_1306	0.13	0.03	0.30	0.11	0.02	0.21	0.14	0.00	0.00	0.00	0.04
8	25945744	T	ALU	ALU_umary_ALU_6641	0.03	0.14	0.00	0.00	0.00	0.03	0.14	0.01	0.00	0.00	0.14
11	24015236	T	ALU	ALU_umary_ALU_8446	0.04	0.14	0.00	0.00	0.00	0.04	0.14	0.01	0.00	0.00	0.14
7	20170340	T	L1	L1_umary_LINE1_1463	0.06	0.18	0.03	0.00	0.00	0.10	0.14	0.01	0.00	0.00	0.18
6	88698077	T	ALU	ALU_umary_ALU_5356	0.13	0.28	0.01	0.13	0.06	0.15	0.14	0.00	0.02	0.00	0.06
8	37037400	T	ALU	ALU_umary_ALU_6683	0.05	0.16	0.00	0.01	0.02	0.06	0.14	0.00	0.00	0.00	0.14
2	242120145	C	ALU	ALU_umary_ALU_2085	0.04	0.14	0.00	0.00	0.00	0.05	0.14	0.00	0.00	0.00	0.14
7	21624579	G	ALU	ALU_umary_ALU_5882	0.04	0.14	0.00	0.00	0.00	0.04	0.14	0.01	0.00	0.00	0.14
10	25957326	G	ALU	ALU_umary_ALU_7854	0.27	0.44	0.10	0.28	0.16	0.35	0.14	0.02	0.00	0.00	0.05
3	145523377	C	ALU	ALU_umary_ALU_2776	0.03	0.14	0.00	0.00	0.00	0.03	0.14	0.01	0.00	0.00	0.14
12	18682072	T	ALU	ALU_umary_ALU_9039	0.04	0.16	0.00	0.01	0.00	0.06	0.14	0.00	0.00	0.00	0.13
5	3666566	T	ALU	ALU_umary_ALU_4038	0.07	0.20	0.00	0.05	0.00	0.08	0.14	0.01	0.00	0.01	0.10
2	170095035	T	ALU	ALU_umary_ALU_1741	0.04	0.15	0.00	0.00	0.02	0.05	0.14	0.01	0.00	0.00	0.13
10	32194794	A	SVA	SVA_umary_SVA_443	0.04	0.14	0.00	0.00	0.00	0.04	0.14	0.00	0.00	0.00	0.14
12	6066803	A	ALU	ALU_umary_ALU_8967	0.04	0.15	0.00	0.00	0.00	0.04	0.14	0.01	0.00	0.00	0.13
16	4097485	T	ALU	ALU_umary_ALU_10936	0.04	0.14	0.00	0.00	0.00	0.05	0.14	0.01	0.00	0.00	0.14
13	25567827	T	ALU	ALU_umary_ALU_9630	0.11	0.02	0.28	0.12	0.05	0.07	0.14	0.00	0.02	0.00	0.08
1	100938616	T	ALU	ALU_umary_ALU_400	0.04	0.14	0.00	0.00	0.00	0.04	0.14	0.00	0.00	0.00	0.13
4	160628485	T	ALU	ALU_umary_ALU_3860	0.04	0.14	0.00	0.00	0.00	0.05	0.14	0.00	0.00	0.00	0.14
10	36467306	T	ALU	ALU_umary_ALU_7909	0.12	0.26	0.00	0.15	0.04	0.14	0.14	0.00	0.00	0.00	0.04
4	81899822	T	ALU	ALU_umary_ALU_3448	0.11	0.27	0.11	0.02	0.05	0.12	0.14	0.01	0.00	0.01	0.23
6	103330618	T	ALU	ALU_umary_ALU_5437	0.04	0.14	0.00	0.00	0.00	0.04	0.14	0.00	0.00	0.00	0.13
16	70497868	T	ALU	ALU_umary_ALU_11105	0.03	0.14	0.00	0.00	0.00	0.03	0.14	0.01	0.00	0.00	0.14
13	100799106	T	ALU	ALU_umary_ALU_10067	0.04	0.14	0.00	0.00	0.00	0.04	0.14	0.00	0.00	0.00	0.14
8	59645639	T	ALU	ALU_umary_ALU_6762	0.04	0.14	0.00	0.00	0.00	0.04	0.14	0.00	0.00	0.00	0.13
8	16871271	A	ALU	ALU_umary_ALU_6600	0.03	0.14	0.00	0.00	0.00	0.04	0.14	0.00	0.00	0.00	0.13

Table 11 (continued).

13	70264401	C	ALU	ALU_umary_ALU_9904	0.07	0.20	0.00	0.05	0.03	0.09	0.14	0.00	0.01	0.03	0.10
3	89723846	T	ALU	ALU_umary_ALU_2518	0.04	0.14	0.00	0.00	0.00	0.04	0.13	0.00	0.00	0.00	0.13
7	145221347	G	ALU	ALU_umary_ALU_6468	0.04	0.14	0.00	0.00	0.00	0.04	0.13	0.00	0.00	0.00	0.13
12	27511918	T	ALU	ALU_umary_ALU_9098	0.04	0.14	0.00	0.00	0.00	0.05	0.13	0.00	0.00	0.00	0.13
14	24794428	G	ALU	ALU_umary_ALU_10151	0.07	0.21	0.01	0.03	0.01	0.08	0.13	0.00	0.00	0.00	0.13
6	76637541	T	L1	L1_umary_LINE1_1333	0.04	0.00	0.15	0.01	0.01	0.02	0.13	0.00	0.00	0.05	0.01
13	25402746	T	ALU	ALU_umary_ALU_9628	0.04	0.14	0.00	0.00	0.00	0.05	0.13	0.00	0.00	0.00	0.13
21	21720992	A	ALU	ALU_umary_ALU_12298	0.03	0.13	0.00	0.00	0.00	0.03	0.13	0.00	0.00	0.00	0.13
5	120317438	G	ALU	ALU_umary_ALU_4631	0.04	0.13	0.00	0.00	0.00	0.04	0.13	0.00	0.00	0.00	0.13
6	99998541	T	ALU	ALU_umary_ALU_5408	0.04	0.14	0.00	0.00	0.01	0.04	0.13	0.01	0.00	0.00	0.13
1	94100853	T	ALU	ALU_umary_ALU_360	0.03	0.13	0.00	0.00	0.00	0.03	0.13	0.01	0.00	0.00	0.13
18	24334488	T	ALU	ALU_umary_ALU_11587	0.03	0.13	0.00	0.00	0.00	0.04	0.13	0.01	0.00	0.00	0.13
2	100785867	C	ALU	ALU_umary_ALU_1390	0.04	0.13	0.00	0.00	0.00	0.05	0.13	0.00	0.00	0.00	0.13
2	189163017	T	ALU	ALU_umary_ALU_1838	0.10	0.00	0.03	0.18	0.24	0.05	0.13	0.00	0.00	0.00	0.17
22	47239372	T	ALU	ALU_umary_ALU_12524	0.04	0.13	0.00	0.00	0.00	0.05	0.13	0.00	0.00	0.00	0.13
2	230254069	A	ALU	ALU_umary_ALU_2038	0.04	0.13	0.00	0.00	0.00	0.05	0.13	0.00	0.00	0.00	0.13
3	86264950	T	L1	L1_umary_LINE1_634	0.08	0.22	0.04	0.02	0.01	0.09	0.13	0.00	0.02	0.00	0.17
6	136367046	T	ALU	ALU_umary_ALU_5613	0.13	0.01	0.26	0.21	0.07	0.11	0.13	0.00	0.00	0.01	0.20
8	129955767	G	ALU	ALU_umary_ALU_7148	0.04	0.15	0.00	0.00	0.01	0.04	0.13	0.01	0.00	0.00	0.14
2	6517433	C	ALU	ALU_umary_ALU_982	0.18	0.37	0.08	0.15	0.10	0.19	0.13	0.00	0.00	0.00	0.11
9	82082793	C	L1	L1_umary_LINE1_1854	0.06	0.19	0.00	0.04	0.01	0.07	0.13	0.02	0.00	0.01	0.10
18	56307379	T	ALU	ALU_umary_ALU_11732	0.03	0.13	0.00	0.00	0.00	0.03	0.13	0.01	0.00	0.00	0.13
17	12795719	G	ALU	ALU_umary_ALU_11214	0.03	0.13	0.00	0.00	0.00	0.03	0.13	0.00	0.00	0.00	0.13
8	117226394	G	ALU	ALU_umary_ALU_7083	0.04	0.13	0.00	0.00	0.00	0.04	0.13	0.01	0.00	0.00	0.13
2	231470361	G	ALU	ALU_umary_ALU_2043	0.03	0.13	0.00	0.00	0.00	0.03	0.13	0.00	0.00	0.00	0.13
4	150882725	G	ALU	ALU_umary_ALU_3813	0.03	0.13	0.00	0.00	0.00	0.03	0.13	0.00	0.00	0.00	0.13
15	65786350	C	SVA	SVA_umary_SVA_639	0.03	0.13	0.00	0.00	0.00	0.04	0.13	0.01	0.00	0.00	0.13
22	47416411	G	ALU	ALU_umary_ALU_12527	0.04	0.13	0.00	0.00	0.00	0.05	0.13	0.00	0.00	0.00	0.13

Table 11 (continued).

10	127299087	G	ALU	ALU_umary_ALU_8303	0.05	0.16	0.01	0.00	0.00	0.06	0.13	0.00	0.01	0.01	0.14
10	130888837	T	ALU	ALU_umary_ALU_8317	0.04	0.15	0.01	0.00	0.01	0.06	0.13	0.00	0.01	0.00	0.14
15	101569683	T	ALU	ALU_umary_ALU_10926	0.03	0.13	0.00	0.00	0.00	0.04	0.13	0.01	0.00	0.00	0.13
18	76194283	A	ALU	ALU_umary_ALU_11850	0.04	0.13	0.00	0.00	0.00	0.05	0.13	0.00	0.00	0.00	0.13
21	16352947	T	ALU	ALU_umary_ALU_12272	0.05	0.17	0.00	0.03	0.01	0.05	0.13	0.00	0.00	0.00	0.11
12	55689859	C	ALU	ALU_umary_ALU_9224	0.03	0.13	0.00	0.00	0.00	0.04	0.13	0.01	0.00	0.00	0.13
13	92825238	C	ALU	ALU_umary_ALU_10029	0.26	0.06	0.37	0.25	0.42	0.19	0.13	0.03	0.01	0.00	0.12
9	71190135	T	L1	L1_umary_LINE1_1831	0.03	0.13	0.00	0.00	0.00	0.03	0.13	0.00	0.00	0.00	0.13
15	51189312	A	L1	L1_umary_LINE1_2647	0.14	0.04	0.32	0.16	0.07	0.11	0.13	0.00	0.00	0.00	0.07
4	106550763	T	ALU	ALU_umary_ALU_3575	0.10	0.26	0.06	0.05	0.04	0.11	0.13	0.01	0.00	0.01	0.16
1	217826976	C	ALU	ALU_umary_ALU_794	0.04	0.13	0.00	0.00	0.00	0.04	0.13	0.01	0.00	0.01	0.13
9	83859674	T	ALU	ALU_umary_ALU_7523	0.05	0.16	0.00	0.02	0.00	0.07	0.13	0.00	0.00	0.00	0.11
7	147704401	G	ALU	ALU_umary_ALU_6483	0.06	0.00	0.19	0.04	0.02	0.04	0.13	0.00	0.00	0.00	0.03
18	6850865	T	ALU	ALU_umary_ALU_11518	0.03	0.13	0.00	0.00	0.00	0.04	0.13	0.01	0.00	0.00	0.13
8	96473928	T	ALU	ALU_umary_ALU_6962	0.05	0.16	0.00	0.02	0.01	0.08	0.13	0.01	0.00	0.00	0.11
8	106432077	A	ALU	ALU_umary_ALU_7005	0.03	0.13	0.00	0.00	0.00	0.03	0.13	0.00	0.00	0.00	0.13
2	163606874	G	ALU	ALU_umary_ALU_1707	0.04	0.13	0.00	0.00	0.01	0.04	0.13	0.00	0.00	0.00	0.13
19	22543111	T	ALU	ALU_umary_ALU_11899	0.04	0.13	0.00	0.00	0.00	0.05	0.13	0.02	0.00	0.00	0.12
12	37870151	T	ALU	ALU_umary_ALU_9139	0.03	0.13	0.00	0.00	0.00	0.04	0.13	0.03	0.00	0.00	0.13
1	18271115	T	ALU	ALU_umary_ALU_50	0.04	0.13	0.00	0.00	0.00	0.05	0.13	0.00	0.00	0.00	0.13
8	15149782	T	ALU	ALU_umary_ALU_6592	0.03	0.13	0.00	0.00	0.00	0.04	0.13	0.01	0.00	0.00	0.13
3	76551228	A	ALU	ALU_umary_ALU_2458	0.03	0.13	0.00	0.00	0.00	0.04	0.13	0.00	0.00	0.00	0.13
4	69132572	C	SVA	SVA_umary_SVA_214	0.03	0.13	0.00	0.00	0.00	0.03	0.13	0.02	0.00	0.00	0.13
10	21192833	G	ALU	ALU_umary_ALU_7830	0.17	0.34	0.13	0.08	0.05	0.23	0.13	0.00	0.01	0.01	0.19
14	82223469	G	ALU	ALU_umary_ALU_10437	0.08	0.22	0.08	0.00	0.02	0.07	0.13	0.00	0.01	0.00	0.21
5	64611524	C	ALU	ALU_umary_ALU_4338	0.03	0.13	0.00	0.00	0.00	0.03	0.13	0.01	0.00	0.00	0.13
1	86058693	T	ALU	ALU_umary_ALU_322	0.03	0.13	0.00	0.00	0.00	0.04	0.13	0.02	0.00	0.00	0.13
1	162080099	T	ALU	ALU_umary_ALU_544	0.12	0.29	0.14	0.03	0.03	0.10	0.12	0.00	0.00	0.00	0.22

Table 11 (continued).

12	60118795	T	ALU	ALU_umary_ALU_9249	0.03	0.13	0.00	0.00	0.00	0.04	0.12	0.00	0.00	0.00	0.12
6	165427429	C	ALU	ALU_umary_ALU_5751	0.03	0.12	0.00	0.00	0.00	0.04	0.12	0.02	0.00	0.00	0.12
10	34235115	T	ALU	ALU_umary_ALU_7899	0.03	0.12	0.00	0.00	0.00	0.04	0.12	0.00	0.00	0.00	0.12
3	64986947	A	ALU	ALU_umary_ALU_2408	0.03	0.12	0.00	0.00	0.00	0.03	0.12	0.00	0.00	0.00	0.12
6	1133041	T	ALU	ALU_umary_ALU_4917	0.03	0.13	0.00	0.00	0.00	0.04	0.12	0.01	0.00	0.00	0.12
10	24369787	T	ALU	ALU_umary_ALU_7846	0.03	0.13	0.00	0.00	0.00	0.03	0.12	0.01	0.00	0.00	0.12
16	68974502	T	ALU	ALU_umary_ALU_11100	0.03	0.12	0.00	0.00	0.00	0.02	0.12	0.00	0.00	0.00	0.12
1	102240167	T	ALU	ALU_umary_ALU_411	0.03	0.12	0.00	0.00	0.00	0.04	0.12	0.00	0.00	0.00	0.12
3	85663369	T	L1	L1_umary_LINE1_630	0.03	0.13	0.00	0.00	0.00	0.03	0.12	0.00	0.00	0.00	0.12
6	112395529	C	ALU	ALU_umary_ALU_5481	0.03	0.13	0.00	0.00	0.00	0.03	0.12	0.01	0.00	0.00	0.12
5	10537250	G	ALU	ALU_umary_ALU_4080	0.03	0.12	0.00	0.00	0.00	0.04	0.12	0.01	0.00	0.00	0.12
4	55993120	T	ALU	ALU_umary_ALU_3292	0.03	0.12	0.00	0.00	0.00	0.05	0.12	0.01	0.00	0.00	0.12
2	128638904	T	ALU	ALU_umary_ALU_1516	0.19	0.01	0.16	0.25	0.40	0.15	0.12	0.02	0.00	0.01	0.23
8	82425047	T	ALU	ALU_umary_ALU_6880	0.03	0.12	0.00	0.00	0.00	0.04	0.12	0.01	0.00	0.00	0.12
3	166092234	A	L1	L1_umary_LINE1_735	0.03	0.12	0.00	0.00	0.00	0.04	0.12	0.00	0.00	0.00	0.12
4	86229689	T	ALU	ALU_umary_ALU_3473	0.03	0.12	0.00	0.00	0.00	0.03	0.12	0.00	0.00	0.00	0.12
4	79511942	T	ALU	ALU_umary_ALU_3435	0.03	0.12	0.00	0.00	0.00	0.03	0.12	0.01	0.00	0.00	0.12
18	54908055	T	ALU	ALU_umary_ALU_11726	0.03	0.12	0.00	0.00	0.00	0.04	0.12	0.00	0.00	0.00	0.12
14	47982138	T	ALU	ALU_umary_ALU_10285	0.03	0.12	0.00	0.00	0.00	0.04	0.12	0.01	0.00	0.00	0.12
14	43740637	C	ALU	ALU_umary_ALU_10257	0.07	0.00	0.18	0.05	0.01	0.13	0.12	0.01	0.01	0.00	0.04
11	107354635	C	ALU	ALU_umary_ALU_8827	0.03	0.12	0.00	0.00	0.00	0.05	0.12	0.00	0.00	0.00	0.12
3	21800882	C	L1	L1_umary_LINE1_561	0.03	0.12	0.00	0.00	0.00	0.04	0.12	0.01	0.00	0.00	0.12
9	75996063	T	ALU	ALU_umary_ALU_7474	0.03	0.12	0.00	0.00	0.00	0.03	0.12	0.02	0.00	0.00	0.12
2	103950685	T	ALU	ALU_umary_ALU_1399	0.03	0.12	0.00	0.00	0.00	0.03	0.12	0.00	0.00	0.00	0.12
7	85396914	T	ALU	ALU_umary_ALU_6168	0.03	0.12	0.00	0.00	0.00	0.04	0.12	0.00	0.00	0.00	0.12
4	176402062	G	ALU	ALU_umary_ALU_3958	0.03	0.13	0.00	0.00	0.00	0.04	0.12	0.01	0.00	0.00	0.12
3	101597726	C	ALU	ALU_umary_ALU_2573	0.03	0.12	0.00	0.00	0.00	0.03	0.12	0.00	0.00	0.00	0.12
10	10615756	A	ALU	ALU_umary_ALU_7783	0.03	0.12	0.00	0.00	0.00	0.04	0.12	0.00	0.00	0.00	0.12

Table 11 (continued).

1	84070086	A	ALU	ALU_umary_ALU_307	0.03	0.12	0.00	0.00	0.00	0.05	0.12	0.00	0.00	0.00	0.12
15	66666669	G	ALU	ALU_umary_ALU_10779	0.03	0.12	0.00	0.00	0.00	0.03	0.12	0.00	0.00	0.00	0.12
7	24052810	T	ALU	ALU_umary_ALU_5894	0.03	0.12	0.00	0.00	0.00	0.04	0.12	0.00	0.00	0.00	0.12
6	122157421	T	SVA	SVA_umary_SVA_305	0.15	0.32	0.21	0.05	0.05	0.13	0.12	0.01	0.01	0.00	0.22
7	125861744	G	ALU	ALU_umary_ALU_6372	0.10	0.26	0.07	0.05	0.02	0.11	0.12	0.01	0.00	0.00	0.15
11	73276648	G	ALU	ALU_umary_ALU_8637	0.03	0.12	0.00	0.00	0.00	0.03	0.12	0.02	0.00	0.00	0.12
18	76783844	T	ALU	ALU_umary_ALU_11852	0.03	0.12	0.00	0.00	0.00	0.04	0.12	0.00	0.00	0.00	0.12
5	106736020	T	ALU	ALU_umary_ALU_4553	0.03	0.13	0.00	0.00	0.00	0.03	0.12	0.00	0.00	0.00	0.12
5	104650875	T	ALU	ALU_umary_ALU_4539	0.11	0.27	0.07	0.06	0.03	0.12	0.12	0.00	0.00	0.00	0.15
3	34722849	T	ALU	ALU_umary_ALU_2275	0.03	0.12	0.00	0.00	0.00	0.04	0.12	0.00	0.00	0.00	0.12
2	220195382	T	ALU	ALU_umary_ALU_1995	0.03	0.12	0.00	0.00	0.00	0.03	0.12	0.02	0.00	0.00	0.12
11	88240142	C	ALU	ALU_umary_ALU_8713	0.03	0.12	0.00	0.00	0.00	0.05	0.12	0.01	0.00	0.00	0.12
5	39505871	T	ALU	ALU_umary_ALU_4226	0.03	0.12	0.00	0.00	0.00	0.04	0.12	0.00	0.00	0.00	0.12
9	105494179	T	ALU	ALU_umary_ALU_7616	0.05	0.15	0.00	0.03	0.00	0.05	0.12	0.00	0.00	0.01	0.10
6	94263995	A	L1	L1_umary_LINE1_1358	0.03	0.12	0.00	0.00	0.00	0.04	0.12	0.00	0.00	0.00	0.11
13	93810884	T	ALU	ALU_umary_ALU_10035	0.03	0.12	0.00	0.00	0.00	0.04	0.12	0.01	0.00	0.00	0.11
5	162166934	T	ALU	ALU_umary_ALU_4825	0.03	0.12	0.00	0.00	0.00	0.03	0.12	0.00	0.00	0.00	0.11
2	129573988	A	L1	L1_umary_LINE1_398	0.19	0.00	0.23	0.20	0.34	0.18	0.12	0.01	0.01	0.01	0.19
2	56373531	G	ALU	ALU_umary_ALU_1216	0.16	0.11	0.38	0.15	0.07	0.09	0.12	0.00	0.00	0.01	0.01
8	140593332	T	ALU	ALU_umary_ALU_7196	0.09	0.22	0.01	0.09	0.02	0.11	0.11	0.01	0.00	0.00	0.06
18	28732918	T	L1	L1_umary_LINE1_2803	0.21	0.16	0.43	0.18	0.10	0.20	0.11	0.00	0.00	0.00	0.00
2	106452585	T	ALU	ALU_umary_ALU_1406	0.25	0.45	0.19	0.18	0.11	0.30	0.11	0.01	0.00	0.02	0.15
6	123853946	T	L1	L1_umary_LINE1_1395	0.09	0.00	0.02	0.14	0.24	0.04	0.10	0.00	0.00	0.01	0.14
12	41823605	T	ALU	ALU_umary_ALU_9168	0.10	0.20	0.14	0.00	0.01	0.13	0.10	0.00	0.00	0.00	0.20
14	63226292	T	ALU	ALU_umary_ALU_10360	0.30	0.09	0.33	0.34	0.48	0.26	0.10	0.00	0.00	0.01	0.16
11	66008262	T	ALU	ALU_umary_ALU_8623	0.08	0.19	0.13	0.00	0.00	0.05	0.09	0.00	0.00	0.00	0.18
1	171358314	T	ALU	ALU_umary_ALU_592	0.08	0.02	0.01	0.10	0.22	0.05	0.06	0.00	0.00	0.00	0.06
9	119496835	T	ALU	ALU_umary_ALU_7679	0.04	0.00	0.01	0.00	0.17	0.00	0.01	0.00	0.00	0.00	0.00

APPENDIX C.

SUPPLEMENTARY INFORMATION FOR CHAPTER 4

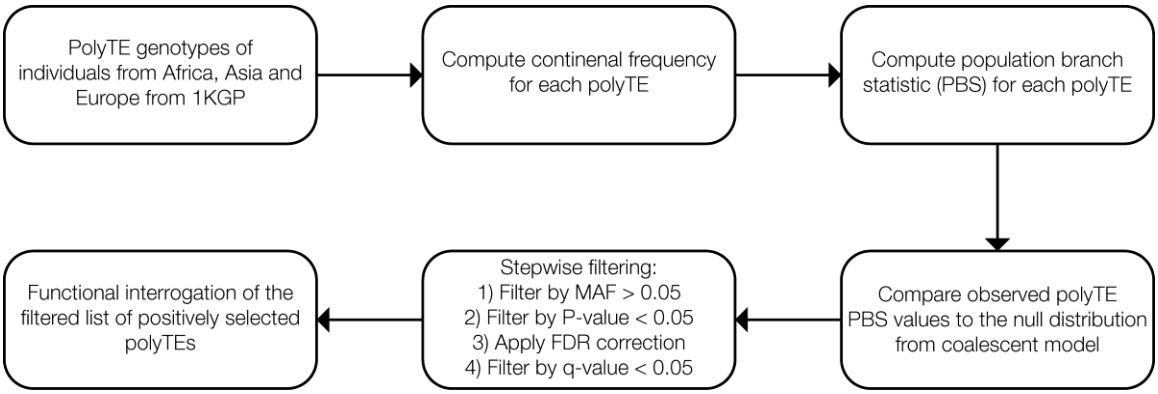

Figure 31 Scheme of the analytical design used in this study.

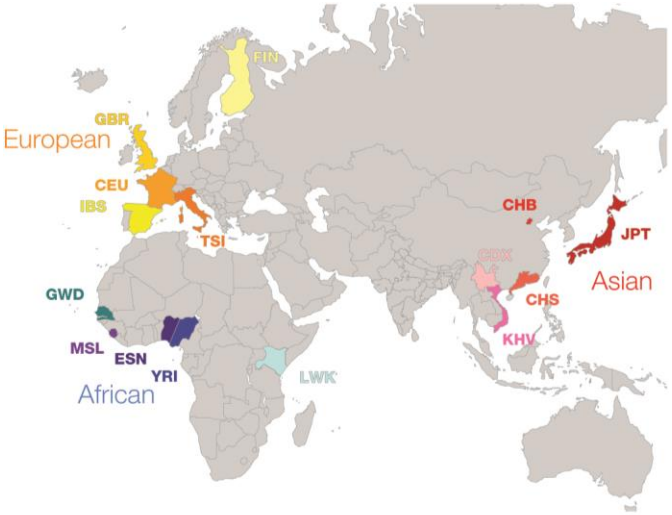

Figure 32 Global populations analyzed in this study.

A total of 14,385 polyTE insertions were characterized for 1,511 individuals from 15 populations belonging to three continental groups: African (blue), Asian (red) and European (gold). PolyTE data from five populations per continental group were analyzed, with the population abbreviations defined in Table 5.

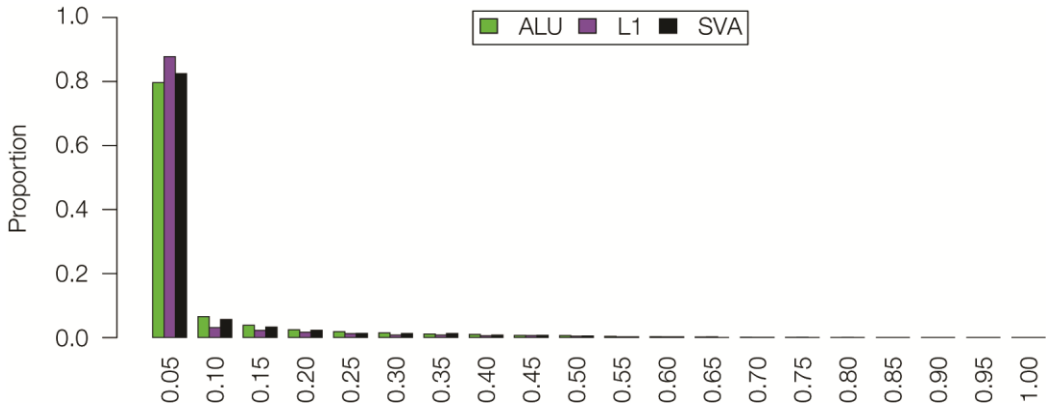

Figure 33 Unfolded allele frequency spectrum for Alu (green), L1 (purple) and SVA polyTE insertions.

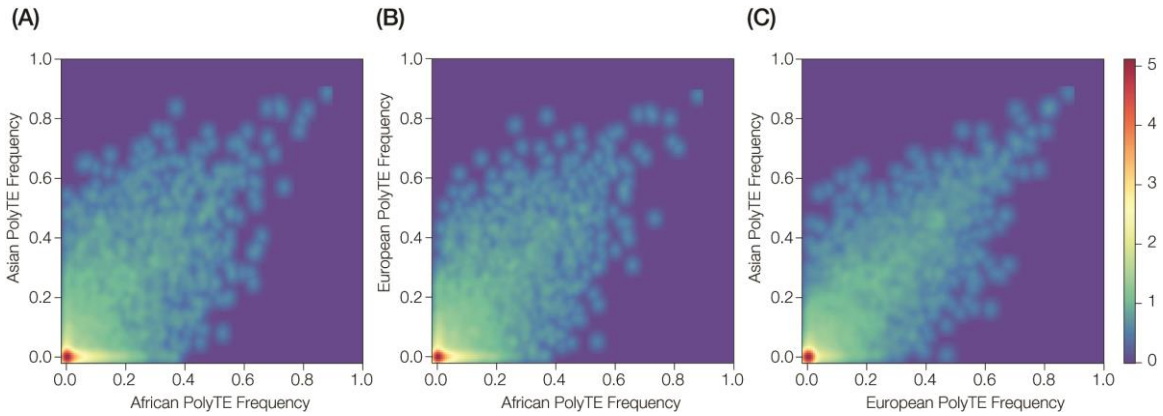

Figure 34 Correlations of polyTE insertion allele frequencies between continental population groups.

The same polyTE insertion allele frequency values as shown in Figure 1 E-G are represented here as heatmaps based on kernel density estimates (KDE).

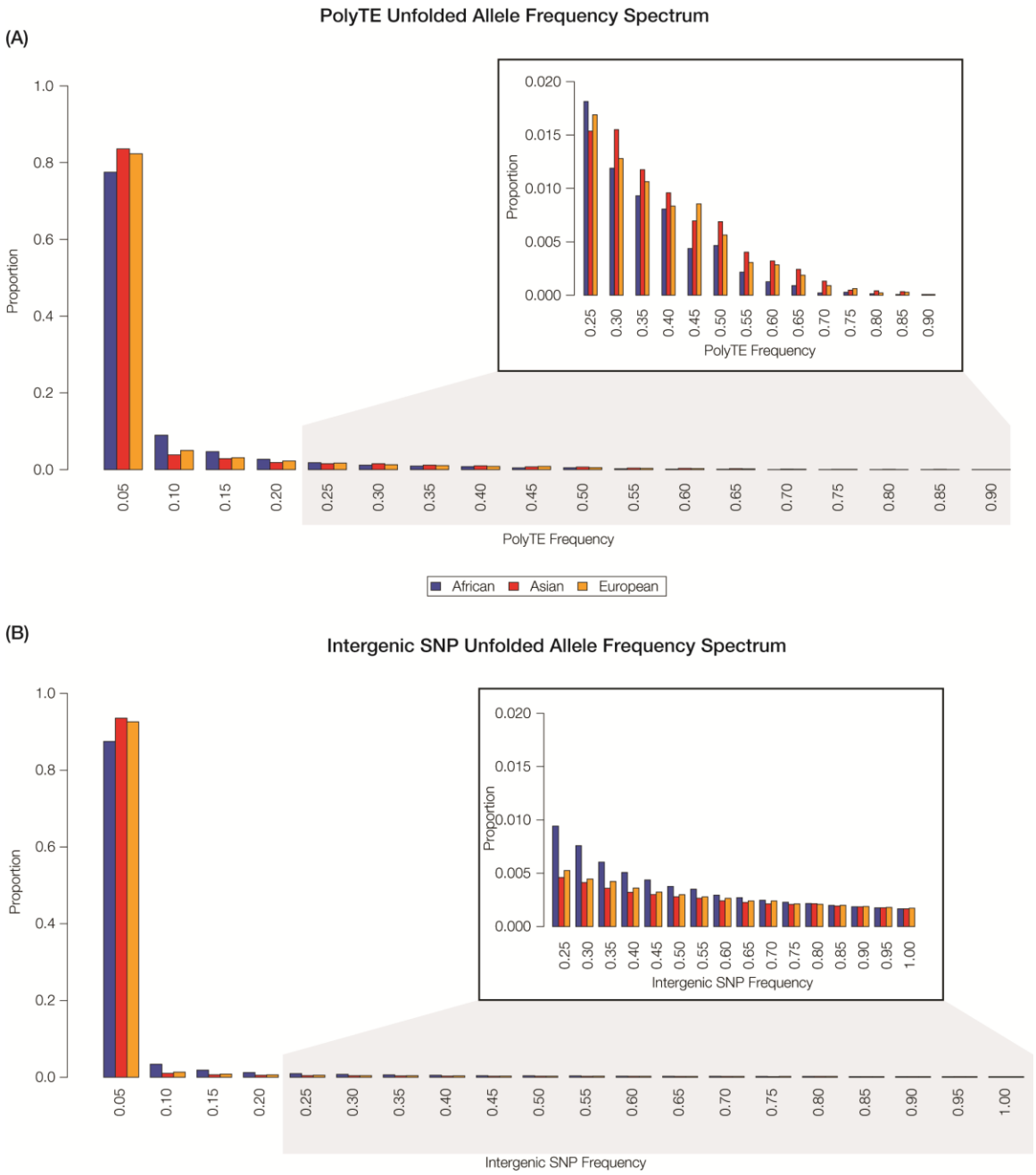

Figure 35 Unfolded allele frequency spectra for polyTE insertions (A) and intergenic SNPs (B) from African (blue), Asian (red) and European (gold) population groups.

The insets expand the higher ranges of the allele frequency spectrum for each plot (≥0.25).

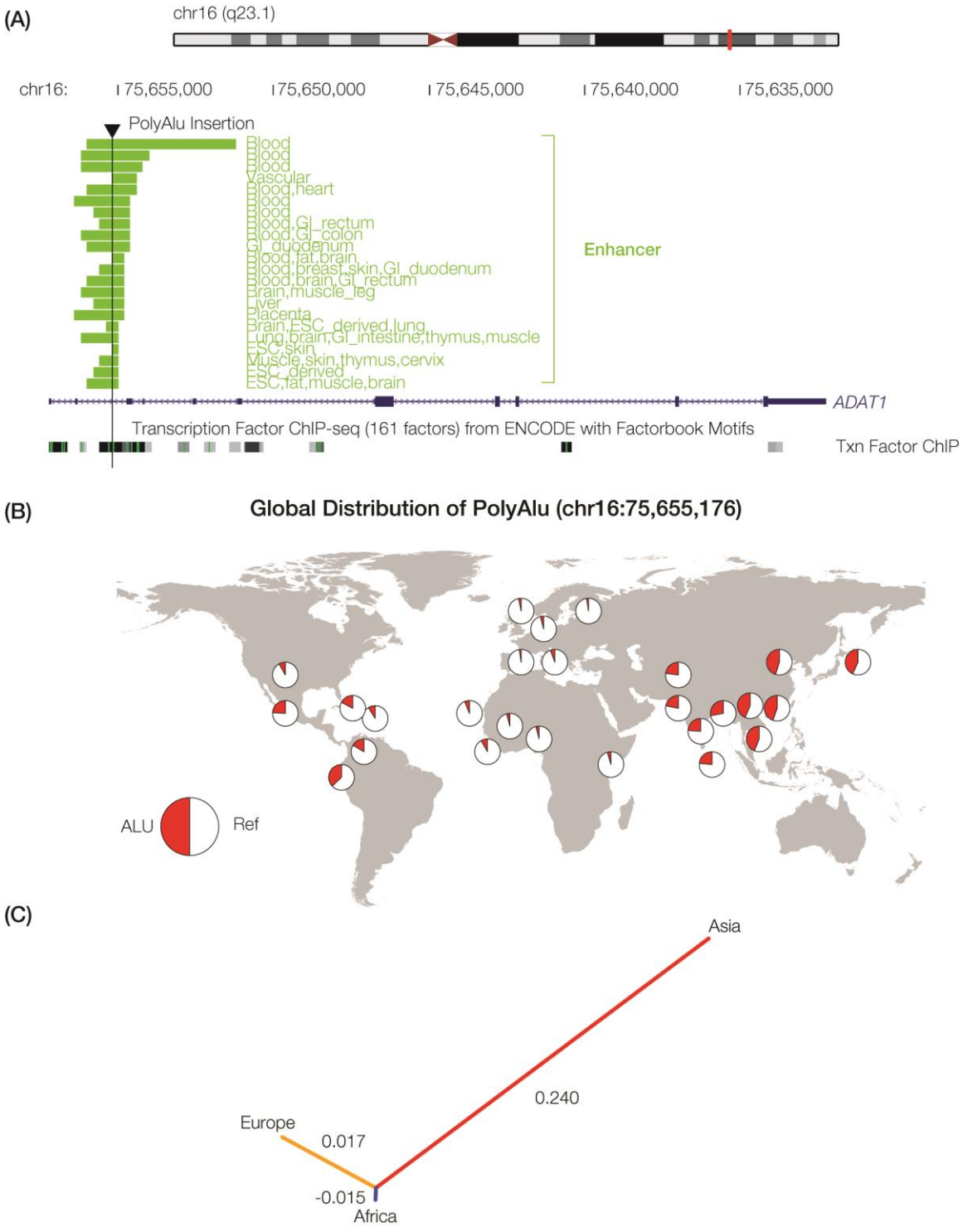

Figure 36 Positively selected polyAlu insertion in the ADAT1 gene.

(A) Chromosome 16 ideogram showing the location (red bar) of the ADAT1 gene on the long arm of chromosome 16. The location of the polyAlu insertion in the ADAT1 gene

model is shown along with co-located enhancer elements, from a number of different tissues, and transcription factor binding sites. (B) Frequencies of the selected polyAlu insertion (red in the pie charts) for the individual populations studied here from Africa, Asia and Europe. (C) Tree with branch lengths scaled to the population group-specific PBS values (shown for each branch).

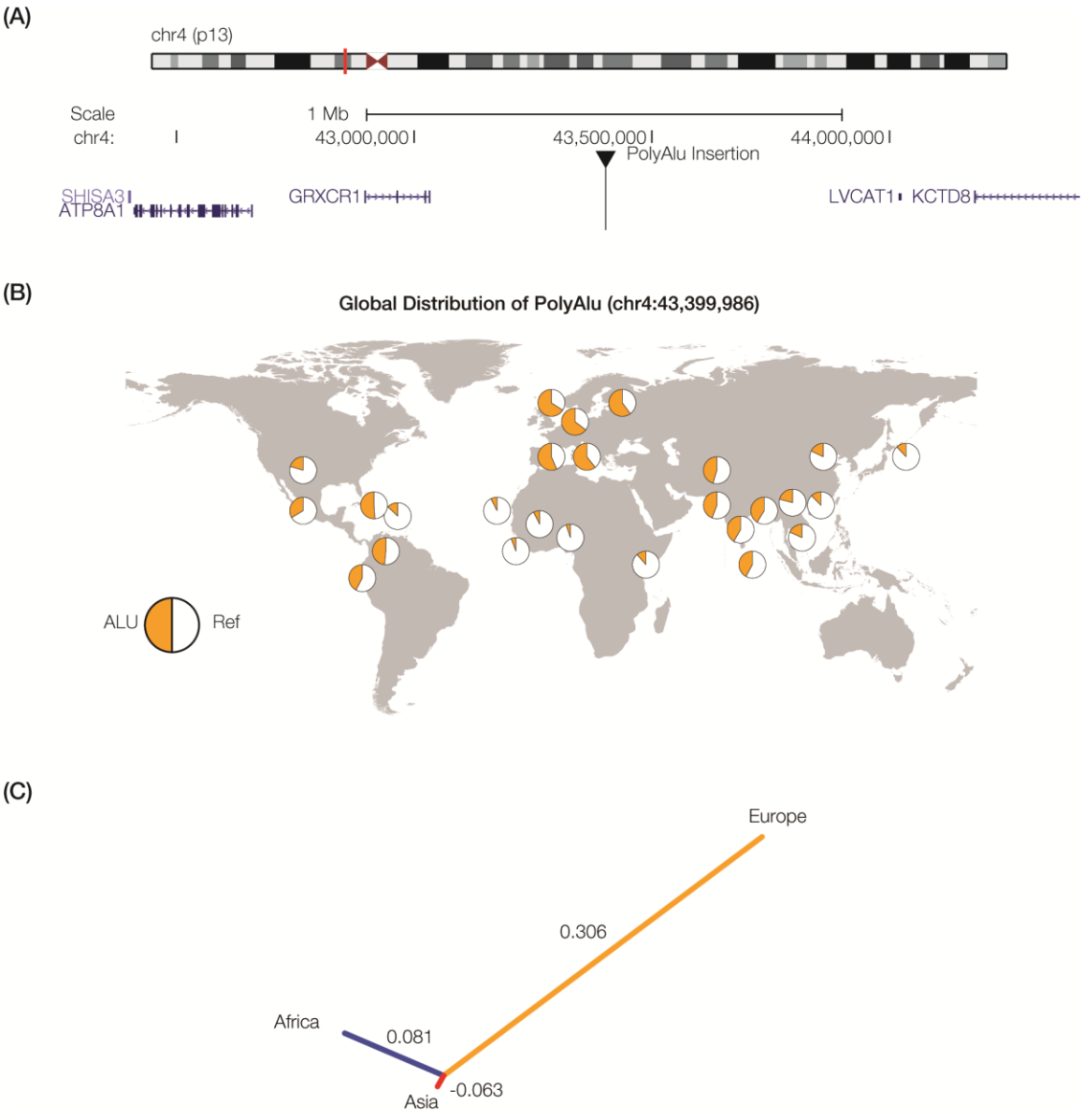

Figure 37 Positively selected polyAlu insertion on chromosome 4.

(A) Chromosome 4 ideogram showing the location (red bar) of the polyAlu insertion on the short arm of chromosome 4. The intergenic location of the polyAlu insertion on chromosome 4 is shown along with the nearby genes. (B) Frequencies of the selected polyAlu insertion (gold in the pie charts) for the individual populations studied here from

Africa, Asia and Europe. (C) Tree with branch lengths scaled to the population group-specific PBS values (shown for each branch).

PUBLICATIONS

1. **Rishishwar, L.,** Wang, L., Wang, J., Yi, S.V., Lachance, J. Jordan, I.K. Polymorphic transposable elements are selected in human populations. *In Prep*.
2. **Rishishwar, L.,** Wang, L., Clayton, E.A., Mariño-Ramírez, L., McDonald, J.F., Jordan, I.K. Population and clinical genetics of human transposable elements in the (post) genomic era. *In Review*.
3. **Rishishwar, L.,** Mariño-Ramírez, L., and Jordan, I.K. (2016). Benchmarking Computational Tools for Polymorphic Transposable Element Detection. Brief Bioinform. doi: 10.1093/bib/bbw072.
4. **Rishishwar, L.,** Tellez Villa, C.E. and Jordan, I.K. (2015). Transposable element polymorphisms recapitulate human evolution. Mob. DNA. 6: 21

5. **Rishishwar, L.** and Jordan, I.K. Implications of human evolution and admixture for mitochondrial replacement therapy. *In Review*.
6. Clayton, E.A., Wang, L., **Rishishwar, L.,** Wang, J., McDonald, J.F., Jordan, I.K. Dynamics of transposable element expression and insertion in cancer. *In review*.
7. Wang, L., **Rishishwar, L.,** Mariño-Ramírez, L., Jordan, I.K. Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. *In Review*.
8. Hu, F., **Rishishwar, L.,** Sivadas, A., Mitchell, G., Jordan, I.K., Murphy, T., Gilsdorf, J., Mayer, L. and Wang, X. (2016). Comparative genomic analysis of

Haemophilus haemolyticus and non-typeable *Haemophilus influenzae* and a new testing scheme for their discrimination. J. Clin. Micro. *In Press.*

9. Gupta, A., Jordan, I.K., **Rishishwar, L.** (2016). stringMLST: a fast k-mer based tool for multi locus sequence typing. Bioinformatics. *In Press.*
10. Orata, F.D., Xu, Y., Gladney, L.M., **Rishishwar, L.,** Case, R.J., Boucher, Y., Jordan, I.K., Tarr, C.L. (2016). Characterization of clinical and environmental isolates of Vibrio cidicii sp. nov., a close relative of *Vibrio navarrensis*. Int J Syst Evol Microbiol. doi: 10.1099/ijsem.0.001327.
11. **Rishishwar, L.,** Kraft, C.S., and Jordan, I.K. (2016). Population genomics of reduced vancomycin susceptibility in Staphylococcus aureus. mSphere. 1: e00094-16
12. Hemme, C.L., Green, S.J., **Rishishwar, L.,** Prakash, O., Pettenato, A., Chakraborty, R., Deutchbauer, A.M., Van Nostrand, J.D., Wu, L., He, L., Jordan, I.K., Hazen, T.C., Arkin, A.P., Kostka, J.E. and Zhou, J. (2016). Lateral gene transfer in a heavy metal-contaminated groundwater microbial community. mBio. 7: e02234-15
13. Medina Rivas, M.A., Norris, E.T., **Rishishwar, L.,** Conley, A.B., Medrano Trochez, C., Valderrama-Aguirre, A., Vannberg, F.O., Mariño-Ramírez, L., and Jordan, I.K. (2016). El Chocó Colombia: a hotspot of human biodiversity. Rev. Biodivers. Neotrop. 6: 45-54
14. **Rishishwar, L.,** Conley, A.B., Vidakovic, B. and Jordan, I.K. (2015). A combined evidence Bayesian method for human ancestry inference applied to Afro-Colombians. Gene. 574: 345-351

15. **Rishishwar, L.,** Conley, A.B., Wigington, C.H., Wang, L., Valderrama-Aguirre, A. and Jordan, I.K. (2015). Ancestry, admixture and fitness in Colombian genomes. Sci. Rep. 5: 12376

16. Gladney, L.M., Katz, L., Knipe, K.M., Rowe, L., Conley, A.B., **Rishishwar, L.,** Mariño-Ramírez, L., Jordan, I.K., Tarr, C.L. (2014). Genome Sequences of *Vibrio navarrensis*, a Potential Human Pathogen. Genome Announcements. 2(6): e01188-14.

17. **Rishishwar, L.,** Petit, R., Kraft, C.S., Jordan, I.K. (2014). A genome sequence based discriminator for vancomycin-intermediate *Staphylococcus aureus*. J Bacteriol. 196(5): 940-948

18. Sebastian, A., **Rishishwar, L.,** Wang, J., Bernard, K.F., Conley, A.B., McCarty, N.A., Jordan, I.K. (2013). Origin and evolution of the cystic fibrosis transmembrane regulator protein R domain. Gene. 523(2):137-46

19. **Rishishwar, L.,** Katz, L.S., Sharma, N.V., Rowe, L., Frace, M., Dolan Thomas, J., Harcourt, B.H., Mayer, L.W., Jordan, I.K. (2012). The genomic basis of a polyagglutinating isolate of *Neisseria meningitidis*. J Bacteriol. 194(20): 5649-5656.

20. **Rishishwar, L.,** Varghese, N., Tyagi, E., Harvey, S.C., Jordan, I.K., McCarty, N.A. (2012). Relating the disease mutation spectrum to the evolution of the cystic fibrosis transmembrane conductance regulator (CFTR). PLoS One, 7(8): e42336.

21. Kostka, J.E., Green, S.J., **Rishishwar, L.,** Prakash, O., Katz, L.S., Marino-Ramirez, L., Jordan, I.K., Watson, D.B., Brown, S.D., Palumbo, A.V., Brooks, S.C. (2012). Genome sequences for six Rhodanobacter strains isolated from soils and the

terrestrial subsurface with variable denitrification capabilities. J Bacteriol. 194(16): 4461.

22. **Rishishwar, L.,** Conley, A.B., Rogowski, K.K., Jiménez-Madrid, J.H., Valderrama-Aguirre, A., Jordan, I.K., Vannberg, F.V. Differential diagnosis of syndromic versus non-syndromic developmental delay in a Colombian patient using whole genome sequence analysis. *In Prep*.
23. Norris, E.T., **Rishishwar, L.,** Pentz, J.T., Wang, X., Mayer, L.W., Jordan, I.K. The genomic basis of capsule switching in the Hajj clone of *Neisseria meningitidis. In Prep*.
24. Chande, A.T., **Rishishwar, L.,** Watve, S., Jordan, I.K., Hammer, B.K. Characterizing novel T6SS in environmental isolates of *Vibrio cholerae*. *In Prep*.
25. Watve, S., Chande, A.T., **Rishishwar, L.,** Bernardy, E.E., Mariño-Ramírez, L., Jordan, I.K., Hammer, B.K. Whole genome sequence of 24 environmental *Vibrio cholerae* isolates. *In Prep*.
26. Sepúlveda-Torres, L.d.C., **Rishishwar, L.,** Rogers, M.L., Ríos-Olivares, E., Jordan, I.K., Boukli, N., Cubano, L.A. A Decade of Drug Resistance and Associated Mutations in a Population of HIV-1+ Puerto Ricans: 2002 – 2011. *In Prep*.

REFERENCES

1. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

2. de Koning, A.P., et al., *Repetitive elements may comprise over two-thirds of the human genome.* PLoS Genet, 2011. **7**(12): p. e1002384.

3. Ray, D.A. and M.A. Batzer, *Reading TE leaves: new approaches to the identification of transposable element insertions.* Genome Res, 2011. **21**(6): p. 813-20.

4. Mc, C.B., *The origin and behavior of mutable loci in maize.* Proc Natl Acad Sci U S A, 1950. **36**(6): p. 344-55.

5. Levin, H.L. and J.V. Moran, *Dynamic interactions between transposable elements and their hosts.* Nat Rev Genet, 2011. **12**(9): p. 615-27.

6. Sijen, T. and R.H. Plasterk, *Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi.* Nature, 2003. **426**(6964): p. 310-4.

7. Consortium, C.e.S., *Genome sequence of the nematode C. elegans: a platform for investigating biology.* Science, 1998. **282**(5396): p. 2012-8.

8. Schnable, P.S., et al., *The B73 maize genome: complexity, diversity, and dynamics.* Science, 2009. **326**(5956): p. 1112-5.

9. Kidwell, M.G., *Transposable elements and the evolution of genome size in eukaryotes.* Genetica, 2002. **115**(1): p. 49-63.

10. Petrov, D.A., *Mutational equilibrium model of genome size evolution.* Theor Popul Biol, 2002. **61**(4): p. 531-44.

11. Kapitonov, V.V. and J. Jurka, *A universal classification of eukaryotic transposable elements implemented in Repbase.* Nat Rev Genet, 2008. **9**(5): p. 411-2; author reply 414.

12. Feschotte, C. and E.J. Pritham, *DNA transposons and the evolution of eukaryotic genomes.* Annu Rev Genet, 2007. **41**: p. 331-68.

13. Kazazian, H.H., Jr., *Mobile elements: drivers of genome evolution.* Science, 2004. **303**(5664): p. 1626-32.

14. Smit, A.F., *The origin of interspersed repeats in the human genome.* Curr Opin Genet Dev, 1996. **6**(6): p. 743-8.

15. Han, J.S., *Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions.* Mob DNA, 2010. **1**(1): p. 15.

16. Smit, A.F., *Interspersed repeats and other mementos of transposable elements in mammalian genomes.* Curr Opin Genet Dev, 1999. **9**(6): p. 657-63.

17. Batzer, M.A. and P.L. Deininger, *Alu repeats and human genomic diversity.* Nat Rev Genet, 2002. **3**(5): p. 370-9.

18. Schmid, C.W., *Does SINE evolution preclude Alu function?* Nucleic Acids Res, 1998. **26**(20): p. 4541-50.

19. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

20. Hackenberg, M., et al., *The biased distribution of Alus in human isochores might be driven by recombination.* J Mol Evol, 2005. **60**(3): p. 365-77.

21. Medstrand, P., L.N. van de Lagemaat, and D.L. Mager, *Retroelement distributions in the human genome: variations associated with age and proximity to genes.* Genome Res, 2002. **12**(10): p. 1483-95.

22. Brookfield, J.F., *Selection on Alu sequences?* Curr Biol, 2001. **11**(22): p. R900-1.

23. Stenger, J.E., et al., *Biased distribution of inverted and direct Alus in the human genome: implications for insertion, exclusion, and genome stability.* Genome Res, 2001. **11**(1): p. 12-27.

24. Lobachev, K.S., et al., *Inverted Alu repeats unstable in yeast are excluded from the human genome.* EMBO J, 2000. **19**(14): p. 3822-30.

25. Huda, A., L. Marino-Ramirez, and I.K. Jordan, *Epigenetic histone modifications of human transposable elements: genome defense versus exaptation.* Mob DNA, 2010. **1**(1): p. 2.

26. Gould, S.J. and E.S. Vrba, *Exaptation—a missing term in the science of form.* Paleobiology, 1982. **8**(01): p. 4-15.

27. Brouha, B., et al., *Hot L1s account for the bulk of retrotransposition in the human population.* Proc Natl Acad Sci U S A, 2003. **100**(9): p. 5280-5.

28. Kazazian, H.H., Jr., et al., *Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man.* Nature, 1988. **332**(6160): p. 164-6.

29. Batzer, M.A. and P.L. Deininger, *A human-specific subfamily of Alu sequences.* Genomics, 1991. **9**(3): p. 481-7.

30. Batzer, M.A., et al., *Amplification dynamics of human-specific (HS) Alu family members.* Nucleic Acids Res, 1991. **19**(13): p. 3619-23.

31. Ostertag, E.M., et al., *SVA elements are nonautonomous retrotransposons that cause disease in humans.* Am J Hum Genet, 2003. **73**(6): p. 1444-51.

32. Wang, H., et al., *SVA elements: a hominid-specific retroposon family.* J Mol Biol, 2005. **354**(4): p. 994-1007.

33. Burton, F.H., et al., *Conservation throughout mammalia and extensive protein-encoding capacity of the highly repeated DNA long interspersed sequence one.* J Mol Biol, 1986. **187**(2): p. 291-304.

34. Fanning, T.G. and M.F. Singer, *LINE-1: a mammalian transposable element.* Biochim Biophys Acta, 1987. **910**(3): p. 203-12.

35. Moran, J.V., et al., *High frequency retrotransposition in cultured mammalian cells.* Cell, 1996. **87**(5): p. 917-27.

36. Dewannieux, M., C. Esnault, and T. Heidmann, *LINE-mediated retrotransposition of marked Alu sequences.* Nat Genet, 2003. **35**(1): p. 41-8.

37. Salem, A.H., et al., *Recently integrated Alu elements and human genomic diversity.* Mol Biol Evol, 2003. **20**(8): p. 1349-61.

38. Schmid, C.W. and P.L. Deininger, *Sequence organization of the human genome.* Cell, 1975. **6**(3): p. 345-58.

39. Ullu, E. and C. Tschudi, *Alu sequences are processed 7SL RNA genes.* Nature, 1984. **312**(5990): p. 171-2.

40. Ono, M., M. Kawakami, and T. Takezawa, *A novel human nonviral retroposon derived from an endogenous retrovirus.* Nucleic Acids Res, 1987. **15**(21): p. 8725-37.

41. Shen, L., et al., *Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication.* J Biol Chem, 1994. **269**(11): p. 8466-76.

42. Sudmant, P.H., et al., *An integrated map of structural variation in 2,504 human genomes.* Nature, 2015. **526**(7571): p. 75-81.

43. Xing, J., et al., *Mobile elements create structural variation: analysis of a complete human genome.* Genome Res, 2009. **19**(9): p. 1516-26.

44. Cordaux, R. and M.A. Batzer, *The impact of retrotransposons on human genome evolution.* Nat Rev Genet, 2009. **10**(10): p. 691-703.

45. Deininger, P.L. and M.A. Batzer, *Alu repeats and human disease.* Mol Genet Metab, 1999. **67**(3): p. 183-93.

46. Perna, N.T., et al., *Alu insertion polymorphism: a new type of marker for human population studies.* Hum Biol, 1992. **64**(5): p. 641-8.

47. Batzer, M.A., et al., *African origin of human-specific polymorphic Alu insertions.* Proc Natl Acad Sci U S A, 1994. **91**(25): p. 12288-92.

48. Stoneking, M., et al., *Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa.* Genome Res, 1997. **7**(11): p. 1061-71.

49. Witherspoon, D.J., et al., *Human population genetic structure and diversity inferred from polymorphic L1(LINE-1) and Alu insertions.* Hum Hered, 2006. **62**(1): p. 30-46.

50. Ray, D.A., et al., *Inference of human geographic origins using Alu insertion polymorphisms.* Forensic Sci Int, 2005. **153**(2-3): p. 117-24.

51. Terreros, M.C., et al., *Insights on human evolution: an analysis of Alu insertion polymorphisms.* J Hum Genet, 2009. **54**(10): p. 603-11.

52. Watkins, W.S., et al., *Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms.* Genome Res, 2003. **13**(7): p. 1607-18.

53. Batzer, M.A., et al., *Genetic variation of recent Alu insertions in human populations.* J Mol Evol, 1996. **42**(1): p. 22-9.

54. Bamshad, M., et al., *Genetic evidence on the origins of Indian caste populations.* Genome Res, 2001. **11**(6): p. 994-1004.

55. Nasidze, I., et al., *Alu insertion polymorphisms and the genetic structure of human populations from the Caucasus.* Eur J Hum Genet, 2001. **9**(4): p. 267-72.

56. Novick, G.E., et al., *Polymorphic Alu insertions and the Asian origin of Native American populations.* Hum Biol, 1998. **70**(1): p. 23-39.

57. Bergman, C.M. and H. Quesneville, *Discovering and detecting transposable elements in genome sequences.* Brief Bioinform, 2007. **8**(6): p. 382-92.

58. Ewing, A.D., *Transposable element detection from whole genome sequence data.* Mob DNA, 2015. **6**: p. 24.

59. Stewart, C., et al., *A comprehensive map of mobile element insertion polymorphisms in humans.* PLoS Genet, 2011. **7**(8): p. e1002236.

60. Robb, S.M., et al., *The use of RelocaTE and unassembled short reads to produce high-resolution snapshots of transposable element generated diversity in rice.* G3 (Bethesda), 2013. **3**(6): p. 949-57.

61. Fiston-Lavier, A.S., et al., *T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data.* Nucleic Acids Res, 2015. **43**(4): p. e22.

62. Keane, T.M., K. Wong, and D.J. Adams, *RetroSeq: transposable element discovery from next-generation sequencing data.* Bioinformatics, 2013. **29**(3): p. 389-90.

63. Vidaud, D., et al., *Haemophilia B due to a de novo insertion of a human-specific Alu subfamily member within the coding region of the factor IX gene.* Eur J Hum Genet, 1993. **1**(1): p. 30-6.

64. Chen, J.M., et al., *Detection of two Alu insertions in the CFTR gene.* J Cyst Fibros, 2008. **7**(1): p. 37-43.

65. Oldridge, M., et al., *De novo alu-element insertions in FGFR2 identify a distinct pathological basis for Apert syndrome.* Am J Hum Genet, 1999. **64**(2): p. 446-61.

66. Lester, T., et al., *X-linked immunodeficiency caused by insertion of Alu repeat sequences.* Journal of Medical Genetics, 1997. **34**: p. 1417-1417.

67. Bratthauer, G.L. and T.G. Fanning, *Active LINE-1 retrotransposons in human testicular cancer.* Oncogene, 1992. **7**(3): p. 507-10.

68. Bratthauer, G.L. and T.G. Fanning, *LINE-1 retrotransposon expression in pediatric germ cell tumors.* Cancer, 1993. **71**(7): p. 2383-6.

69. Bratthauer, G.L., R.D. Cardiff, and T.G. Fanning, *Expression of LINE-1 retrotransposons in human breast cancer.* Cancer, 1994. **73**(9): p. 2333-6.

70. Asch, H.L., et al., *Comparative expression of the LINE-1 p40 protein in human breast carcinomas and normal breast tissues.* Oncol Res, 1996. **8**(6): p. 239-47.

71. Bowen, N.J. and I.K. Jordan, *Exaptation of protein coding sequences from transposable elements.* Genome Dyn, 2007. **3**: p. 147-62.

72. Miller, W.J., et al., *P-element homologous sequences are tandemly repeated in the genome of Drosophila guanche.* Proc Natl Acad Sci U S A, 1992. **89**(9): p. 4018-22.

73. Feschotte, C., *Transposable elements and the evolution of regulatory networks.* Nat Rev Genet, 2008. **9**(5): p. 397-405.

74. Conley, A.B., J. Piriyapongsa, and I.K. Jordan, *Retroviral promoters in the human genome.* Bioinformatics, 2008. **24**(14): p. 1563-7.

75. Jordan, I.K., et al., *Origin of a substantial fraction of human regulatory sequences from transposable elements.* Trends in Genetics, 2003. **19**(2): p. 68-72.

76. Marino-Ramirez, L., et al., *Transposable elements donate lineage-specific regulatory sequences to host genomes.* Cytogenetic and Genome Research, 2005. **110**(1-4): p. 333-341.

77. Bejerano, G., et al., *A distal enhancer and an ultraconserved exon are derived from a novel retroposon.* Nature, 2006. **441**(7089): p. 87-90.

78. Chuong, E.B., N.C. Elde, and C. Feschotte, *Regulatory evolution of innate immunity through co-option of endogenous retroviruses.* Science, 2016. **351**(6277): p. 1083-7.

79. Chuong, E.B., et al., *Endogenous retroviruses function as species-specific enhancer elements in the placenta.* Nat Genet, 2013. **45**(3): p. 325-9.

80. Kunarso, G., et al., *Transposable elements have rewired the core regulatory network of human embryonic stem cells.* Nat Genet, 2010. **42**(7): p. 631-4.

81. Notwell, J.H., et al., *A family of transposable elements co-opted into developmental enhancers in the mouse neocortex.* Nat Commun, 2015. **6**: p. 6644.

82. Conley, A.B. and I.K. Jordan, *Cell type-specific termination of transcription by transposable element sequences.* Mob DNA, 2012. **3**(1): p. 15.

83. Kapusta, A., et al., *Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs.* Plos Genetics, 2013. **9**(4).

84. Piriyapongsa, J., L. Marino-Ramirez, and I.K. Jordan, *Origin and evolution of human microRNAs from transposable elements.* Genetics, 2007. **176**(2): p. 1323-37.

85. Weber, M.J., *Mammalian small nucleolar RNAs are mobile genetic elements.* Plos Genetics, 2006. **2**(12): p. 1984-1997.

86. Jacques, P.E., J. Jeyakani, and G. Bourque, *The majority of primate-specific regulatory sequences are derived from transposable elements.* PLoS Genet, 2013. **9**(5): p. e1003504.

87. Pavlicek, A., et al., *Similar integration but different stability of Alus and LINEs in the human genome.* Gene, 2001. **276**(1-2): p. 39-45.

88. Schmidt, D., et al., *Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages.* Cell, 2012. **148**(1-2): p. 335-48.

89. Sundaram, V., et al., *Widespread contribution of transposable elements to the innovation of gene regulatory networks.* Genome Res, 2014. **24**(12): p. 1963-76.

90. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.

91. Barron, M.G., et al., *Population genomics of transposable elements in Drosophila.* Annu Rev Genet, 2014. **48**: p. 561-81.

92. Gonzalez, J., et al., *Genome-wide patterns of adaptation to temperate environments associated with transposable elements in Drosophila.* PLoS Genet, 2010. **6**(4): p. e1000905.

93. Gonzalez, J. and D.A. Petrov, *The adaptive role of transposable elements in the Drosophila genome.* Gene, 2009. **448**(2): p. 124-33.

94. Gonzalez, J., J.M. Macpherson, and D.A. Petrov, *A recent adaptive transposable element insertion near highly conserved developmental loci in Drosophila melanogaster.* Mol Biol Evol, 2009. **26**(9): p. 1949-61.

95. Gonzalez, J., et al., *High rate of recent transposable element-induced adaptation in Drosophila melanogaster.* PLoS Biol, 2008. **6**(10): p. e251.

96. Rebollo, R., et al., *Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms.* PLoS Genet, 2011. **7**(9): p. e1002301.

97. Kuhn, A., et al., *Linkage disequilibrium and signatures of positive selection around LINE-1 retrotransposons in the human genome.* Proc Natl Acad Sci U S A, 2014. **111**(22): p. 8131-6.

98. Konkel, M.K., et al., *Sequence Analysis and Characterization of Active Human Alu Subfamilies Based on the 1000 Genomes Pilot Project.* Genome Biol Evol, 2015. **7**(9): p. 2608-22.

99. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

100. Hancks, D.C. and H.H. Kazazian, Jr., *Active human retrotransposons: variation and disease.* Curr Opin Genet Dev, 2012. **22**(3): p. 191-203.

101. Beck, C.R., et al., *LINE-1 elements in structural variation and disease.* Annu Rev Genomics Hum Genet, 2011. **12**: p. 187-215.

102. Rishishwar, L., C.E. Tellez Villa, and I.K. Jordan, *Transposable element polymorphisms recapitulate human evolution.* Mob DNA, 2015. **6**: p. 21.

103. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.

104. Initiative, N.P.M. *Trans-Omics for Precision Medicine (TOPMed) Program*. April 2, 2015 [cited 2016 Mar 15]; Available from: http://www.nhlbi.nih.gov/research/resources/nhlbi-precision-medicine-initiative/topmed.

105. Barwell, J., C. Powell, and H. Harrison, *The dawn of genomic medicine: the role of the 100,000 Genomes Project in breast care management.* Breast Cancer Management, 2016. **5**(1): p. 7-11.

106. Alkan, C., B.P. Coe, and E.E. Eichler, *Genome structural variation discovery and genotyping.* Nat Rev Genet, 2011. **12**(5): p. 363-76.

107. Smit, A.F.A., R. Hubley, and P. Green. *RepeatMasker Open-4.0*. 2015; Available from: http://www.repeatmasker.org.

108. Jurka, J., et al., *Repbase Update, a database of eukaryotic repetitive elements.* Cytogenet Genome Res, 2005. **110**(1-4): p. 462-7.

109. Jiang, C., et al., *ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data.* BMC Bioinformatics, 2015. **16**: p. 72.

110. Zhuang, J., et al., *TEMP: a computational method for analyzing transposable element polymorphism in populations.* Nucleic Acids Res, 2014. **42**(11): p. 6826-38.

111. Thung, D.T., et al., *Mobster: accurate detection of mobile element insertions in next generation sequencing data.* Genome Biol, 2014. **15**(10): p. 488.

112. Wu, J., et al., *Tangram: a comprehensive toolbox for mobile element insertion detection.* BMC Genomics, 2014. **15**: p. 795.

113. Bergman, C.M. and H. Quesneville. *TE Tools @ Bergman Lab*. 2007 [cited 2016 Mar 15]; Available from: http://bergmanlab.ls.manchester.ac.uk/?page_id=295.

114. OMICtools. *Transposable element detection tools*. [cited 2016 Mar 15]; Available from: http://omictools.com/transposon-detection-category.

115. Zook, J.M., et al., *Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls.* Nat Biotechnol, 2014. **32**(3): p. 246-51.

116. Genomes Project, C. *A global reference for human genetic variation*. 2015 [cited 2016 Mar 15]; Available from: ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp.

117. Bao, W., K.K. Kojima, and O. Kohany, *Repbase Update, a database of repetitive elements in eukaryotic genomes.* Mob DNA, 2015. **6**: p. 11.

118. Huang, W., et al., *ART: a next-generation sequencing read simulator.* Bioinformatics, 2012. **28**(4): p. 593-4.

119. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

120. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

121. Ray, D.A., et al., *SINEs of a nearly perfect character.* Syst Biol, 2006. **55**(6): p. 928-35.

122. Jakobsson, M., et al., *Genotype, haplotype and copy-number variation in worldwide human populations.* Nature, 2008. **451**(7181): p. 998-1003.

123. Lohmueller, K.E., et al., *Proportionally more deleterious genetic variation in European than in African populations.* Nature, 2008. **451**(7181): p. 994-7.

124. Li, J.Z., et al., *Worldwide human relationships inferred from genome-wide patterns of variation.* Science, 2008. **319**(5866): p. 1100-4.

125. Bryc, K., et al., *Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations.* Proc Natl Acad Sci U S A, 2010. **107 Suppl 2**: p. 8954-61.

126. Zakharia, F., et al., *Characterizing the admixed African ancestry of African Americans.* Genome Biol, 2009. **10**(12): p. R141.

127. Reich, D., et al., *Reconstructing Native American population history.* Nature, 2012. **488**(7411): p. 370-4.

128. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.

129. Ding, L., et al., *Comparison of measures of marker informativeness for ancestry and admixture mapping.* BMC Genomics, 2011. **12**: p. 622.

130. Collins-Schramm, H.E., et al., *Ethnic-difference markers for use in mapping by admixture linkage disequilibrium.* Am J Hum Genet, 2002. **70**(3): p. 737-50.

131. Smith, M.W. and S.J. O'Brien, *Mapping by admixture linkage disequilibrium: advances, limitations and guidelines.* Nat Rev Genet, 2005. **6**(8): p. 623-32.

132. Winkler, C.A., G.W. Nelson, and M.W. Smith, *Admixture mapping comes of age.* Annu Rev Genomics Hum Genet, 2010. **11**: p. 65-89.

133. Weir, B.S. and C.C. Cockerham, *Estimating F-Statistics for the Analysis of Population-Structure.* Evolution, 1984. **38**(6): p. 1358-1370.

134. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-8.

135. Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Mol Biol Evol, 1987. **4**(4): p. 406-25.

136. Tamura, K., et al., *MEGA6: Molecular Evolutionary Genetics Analysis version 6.0.* Mol Biol Evol, 2013. **30**(12): p. 2725-9.

137. Cann, H.M., et al., *A human genome diversity cell line panel.* Science, 2002. **296**(5566): p. 261-2.

138. Howie, B.N., P. Donnelly, and J. Marchini, *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.* PLoS Genet, 2009. **5**(6): p. e1000529.

139. Orgel, L.E. and F.H. Crick, *Selfish DNA: the ultimate parasite.* Nature, 1980. **284**(5757): p. 604-7.

140. Doolittle, W.F. and C. Sapienza, *Selfish genes, the phenotype paradigm and genome evolution.* Nature, 1980. **284**(5757): p. 601-3.

141. Hickey, D.A., *Selfish DNA: a sexually-transmitted nuclear parasite.* Genetics, 1982. **101**(3-4): p. 519-31.

142. Burns, K.H. and J.D. Boeke, *Human transposon tectonics.* Cell, 2012. **149**(4): p. 740-52.

143. Chenais, B., *Transposable elements in cancer and other human diseases.* Curr Cancer Drug Targets, 2015. **15**(3): p. 227-42.

144. Hancks, D.C. and H.H. Kazazian, Jr., *Roles for retrotransposon insertions in human disease.* Mob DNA, 2016. **7**: p. 9.

145. Reilly, M.T., et al., *The role of transposable elements in health and diseases of the central nervous system.* J Neurosci, 2013. **33**(45): p. 17577-86.

146. Solyom, S., et al., *Extensive somatic L1 retrotransposition in colorectal tumors.* Genome Res, 2012. **22**(12): p. 2328-38.

147. Rishishwar, L., L. Marino-Ramirez, and I.K. Jordan, *Benchmarking computational tools for polymorphic transposable element detection.* Brief Bioinform, 2016.

148. Cooper, G.M., et al., *Distribution and intensity of constraint in mammalian genomic sequence.* Genome Res, 2005. **15**(7): p. 901-13.

149. Davydov, E.V., et al., *Identifying a high fraction of the human genome to be under selective constraint using GERP++.* PLoS Comput Biol, 2010. **6**(12): p. e1001025.

150. Gronau, I., et al., *Bayesian inference of ancient human demography from individual genome sequences.* Nat Genet, 2011. **43**(10): p. 1031-4.

151. Yi, X., et al., *Sequencing of 50 human exomes reveals adaptation to high altitude.* Science, 2010. **329**(5987): p. 75-8.

152. Quinlan, A.R., *BEDTools: The Swiss-Army Tool for Genome Feature Analysis.* Curr Protoc Bioinformatics, 2014. **47**: p. 11 12 1-34.

153. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D733-45.

154. Speir, M.L., et al., *The UCSC Genome Browser database: 2016 update.* Nucleic Acids Res, 2016. **44**(D1): p. D717-25.

155. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506-11.

156. t Hoen, P.A., et al., *Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories.* Nat Biotechnol, 2013. **31**(11): p. 1015-22.

157. Lappalainen, T., et al. *Transcriptome and genome sequencing uncovers functional variation in humans*. 2013 [cited 2014 Dec 9th]; Available from: ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-1/analysis_results/.

158. Shabalin, A.A., *Matrix eQTL: ultra fast eQTL analysis via large matrix operations.* Bioinformatics, 2012. **28**(10): p. 1353-8.

159. NHGRI. *The Human Genome Project Completion: Frequently Asked Questions*. 2003 October 30, 2010 October 3, 2016]; Available from: https://www.genome.gov/11006943/.

160. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.

161. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing technologies.* Nat Rev Genet, 2016. **17**(6): p. 333-51.

162. Cancer Genome Atlas Research, N., et al., *The Cancer Genome Atlas Pan-Cancer analysis project.* Nat Genet, 2013. **45**(10): p. 1113-20.

163. Zhang, J., et al., *International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data.* Database (Oxford), 2011. **2011**: p. bar026.

164. Consortium, U.K., et al., *The UK10K project identifies rare variants in health and disease.* Nature, 2015. **526**(7571): p. 82-90.

165. Project Team, S.G., *The Saudi Human Genome Program: An oasis in the desert of Arab medicine is providing clues to genetic disease.* IEEE Pulse, 2015. **6**(6): p. 22-6.

166. Mallick, S., et al., *The Simons Genome Diversity Project: 300 genomes from 142 diverse populations.* Nature, 2016.

167. Pagani, L., et al., *Genomic analyses inform on migration events during the peopling of Eurasia.* Nature, 2016.

168. Zook, J.M., et al., *Extensive sequencing of seven human genomes to characterize benchmark reference materials.* Sci Data, 2016. **3**: p. 160025.

169. Chen, K., et al., *BreakDancer: an algorithm for high-resolution mapping of genomic structural variation.* Nat Methods, 2009. **6**(9): p. 677-81.

170. Hormozdiari, F., et al., *Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes.* Genome Res, 2009. **19**(7): p. 1270-8.

171. Hormozdiari, F., et al., *Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery.* Bioinformatics, 2010. **26**(12): p. i350-7.

172. Xing, J., D.J. Witherspoon, and L.B. Jorde, *Mobile element biology: new possibilities with high-throughput sequencing.* Trends Genet, 2013. **29**(5): p. 280-9.

173. Wheelan, S.J., et al., *Transposon insertion site profiling chip (TIP-chip).* Proc Natl Acad Sci U S A, 2006. **103**(47): p. 17632-7.

174. Witherspoon, D.J., et al., *Mobile element scanning (ME-Scan) by targeted high-throughput sequencing.* BMC Genomics, 2010. **11**: p. 410.

175. Ewing, A.D. and H.H. Kazazian, Jr., *High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes.* Genome Res, 2010. **20**(9): p. 1262-70.

176. Iskow, R.C., et al., *Natural mutagenesis of human genomes by endogenous retrotransposons.* Cell, 2010. **141**(7): p. 1253-61.

177. Beck, C.R., et al., *LINE-1 retrotransposition activity in human genomes.* Cell, 2010. **141**(7): p. 1159-70.

178. Prak, E.T. and H.H. Kazazian, Jr., *Mobile elements and the human genome.* Nat Rev Genet, 2000. **1**(2): p. 134-44.

179. Goldman, M.A., et al., *Replication timing of genes and middle repetitive sequences.* Science, 1984. **224**(4650): p. 686-92.

180. Manuelidis, L. and D.C. Ward, *Chromosomal and nuclear distribution of the HindIII 1.9-kb human DNA repeat segment.* Chromosoma, 1984. **91**(1): p. 28-38.

181. Soriano, P., M. Meunier-Rotival, and G. Bernardi, *The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes.* Proc Natl Acad Sci U S A, 1983. **80**(7): p. 1816-20.

182. Meunier-Rotival, M., et al., *Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA.* Proc Natl Acad Sci U S A, 1982. **79**(2): p. 355-9.

183. Cuny, G., et al., *The major components of the mouse and human genomes. 1. Preparation, basic properties and compositional heterogeneity.* Eur J Biochem, 1981. **115**(2): p. 227-33.

184. Rosenberg, N.A., et al., *Informativeness of genetic markers for inference of ancestry.* Am J Hum Genet, 2003. **73**(6): p. 1402-22.

185. Shedlock, A.M., K. Takahashi, and N. Okada, *SINEs of speciation: tracking lineages with retroposons.* Trends Ecol Evol, 2004. **19**(10): p. 545-53.

186. Risch, N., et al., *Categorization of humans in biomedical research: genes, race and disease.* Genome Biol, 2002. **3**(7): p. comment2007.

187. Collins, F.S., et al., *A vision for the future of genomics research.* Nature, 2003. **422**(6934): p. 835-47.

188. Matyunina, L.V., N.J. Bowen, and J.F. McDonald, *LTR retrotransposons and the evolution of dosage compensation in Drosophila.* BMC Mol Biol, 2008. **9**: p. 55.

189. McDonald, J.F., M.A. Matzke, and A.J. Matzke, *Host defenses to transposable elements and the evolution of genomic imprinting.* Cytogenet Genome Res, 2005. **110**(1-4): p. 242-9.

190. Jordan, I.K., *Evolutionary tinkering with transposable elements.* Proc Natl Acad Sci U S A, 2006. **103**(21): p. 7941-2.

191. Vitti, J.J., S.R. Grossman, and P.C. Sabeti, *Detecting natural selection in genomic data.* Annu Rev Genet, 2013. **47**: p. 97-120.

192. Grossman, S.R., et al., *Identifying recent adaptations in large-scale genomic data.* Cell, 2013. **152**(4): p. 703-13.

193. Sabeti, P.C., et al., *Positive natural selection in the human lineage.* Science, 2006. **312**(5780): p. 1614-20.

194. Rodic, N., et al., *Long interspersed element-1 protein expression is a hallmark of many human cancers.* Am J Pathol, 2014. **184**(5): p. 1280-6.

195. Doucet-O'Hare, T.T., et al., *LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma.* Proc Natl Acad Sci U S A, 2015. **112**(35): p. E4894-900.

196. Doucet-O'Hare, T.T., et al., *Somatically Acquired LINE-1 Insertions in Normal Esophagus Undergo Clonal Expansion in Esophageal Squamous Cell Carcinoma.* Hum Mutat, 2016. **37**(9): p. 942-54.

197. Ewing, A.D., et al., *Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution.* Genome Res, 2015. **25**(10): p. 1536-45.

198. Tubio, J.M., et al., *Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes.* Science, 2014. **345**(6196): p. 1251343.

199. Baillie, J.K., et al., *Somatic retrotransposition alters the genetic landscape of the human brain.* Nature, 2011. **479**(7374): p. 534-7.

200. Shukla, R., et al., *Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma.* Cell, 2013. **153**(1): p. 101-11.

201. Scott, E.C., et al., *A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer.* Genome Res, 2016. **26**(6): p. 745-55.

202. Gibson, G., *Rare and common variants: twenty arguments.* Nat Rev Genet, 2011. **13**(2): p. 135-45.

203. Gymrek, M., et al., *Abundant contribution of short tandem repeats to gene expression variation in humans.* Nat Genet, 2016. **48**(1): p. 22-9.

204. Consortium, G.T., *Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.* Science, 2015. **348**(6235): p. 648-60.